DEPARTMENT OF PSYCHOLOGY
WASHINGTON UNIVERSITY
ST. LOUIS 30, MISSOURI

# THE MEASUREMENT AND EVALUATION OF OVER AND UNDERACHIEVEMENT

A Report of a Conference Held at
Washington University
St. Louis, Missouri

Edited by
Philip H. DuBois
and
Edward V. Hackett



Technical Report No. 8
Office of Naval Research Contract No. Nonr 816(14)
Naval Air Technical Training
April 1963

# THE MEASUREMENT AND EVALUATION OF OVER- AND UNDERACHIEVEMENT

Proceedings of a Conference on

Research Methodology in Training

Prepared under Contract Nonr 816(02)
Between the Office of Naval Research
and Washington University, St. Louis

Edited by:

Philip H. DuBois

and

Edward V. Hackett

Washington University

St. Louis, Missouri

Morning Session, 14 April 1961

Dr. Glenn L. Bryan, presiding

DR. BRYAN: We begin with opening remarks by Dr. Philip H. DuBois.

DR. DuBOIS: This is the fourth conference on Research Methodology in Training which has been sponsored jointly by Washington University and the Office of Naval Research under Contract 816(02). The emphasis in this contract has been on the study of adult learning by procedures which are intermediate between laboratory and psychometric methods. Our special concern today is to examine the nature and measurement of over- and underachievement.

The aim of these conferences has always been to bring together groups, small enough to facilitate free discussions. The role of the discussant is really that of lead-off man. While a discussion period is listed at the end of each session, we expect that there will be questions raised along the way.

DR. BRYAN: The first paper will be presented by Mr. Edward V. Hackett of Washington University. It is entitled, "Some impressions of the Scientific Literature on Over- and Underachievement." Mr. Hackett.

MR. HACKETT: This bibliography on over- and underachievement has been compiled largly from the following sources: Journal of Applied Psychology, Journal of Educational Psychology, Educational and Psychological Measurement, Journal of Counseling Psychology, Personnel Guidance Journal and certain clinical journals. Certain sources were not tapped because of either inaccessibility or inconvenience, especially the British Journal of Educational Psychology and the Japanese Journal of Educational Psychology. While there has been considerable productivity in the area of achievement among the Japanese in recent years, all too often only their English summaries seem to be germane to the topic and to convey any lasting impression. In addition, doctoral and master's theses which treated over- and underachievement were neither sent for nor listed except insofar as they might have found their way into the published literature. Of course, some of the references include unpublished theses in their own bibliographies and the interested reader can follow-up those on his own.

A bibliography of 100 entries is by no means exhaustive due to the plethora of research on this topic in recent years. Some articles were eliminated because they duplicate other authors' work, while others have been by-passed because of vagueness of design or of results or both. Similarly, some articles (33, 41, 42, 78, 83, 96) are included not so much for their own research merit, as for either their historical interest or the extensive bibliography they provide. Some, in fact, are literature reviews.

The impressions one gets from a survey of this topic suggest that much of the work can be classified as an example of underachievement itself. Easton (28) developed a series of hypotheses for differentiating the underachiever from the overachiever which provide an interesting parallel for the examination of these studies.

The parenthetical statement following each hypothesis suggests its application in this context. In general, the underachiever 1) tends to reflect insecurity (most studies seem to end with the familiar "more research is needed in this area"), 2) shows less satisfactory parental relationships (each successive investigator feels compelled to break the umbilical cord of previous research), 3) has more egocentricity (whether with a revised MMPI scale or a new study-habits inventory, each investigator seems to have found the "golden bullet" for differentiating the underachiever from the rest), and 4) manifests less achievement drive (many reports end at the point where they might profitably begin). The analogy is not exact but it points up, to some extent, certain strengths and weaknesses of the research on over- and underachievement in the classroom.

In order that we might end on a positive note, we might first examine some of the problems of achievement research in terms of definition, the criterion and predictors. The first of these centers around definitions. If we arbitrarily name some function in the usual operational manner, then over- and underachievers may be variously defined as those students who: deviate $\pm.5$ standard deviations from the regression line established by the ACE or other ability measure (2, 74); are in the top and bottom 27% or 34% of those who deviate $\pm.5$ standard deviations from the mean (60, 70); deviate $\pm.5$ S.D.'s from the comparison of z-score values for the ACE and grade-point-average (65, 66); who deviate $\pm.67$ sigmas from the predicted average (1); or who exceed or fall short of unity in a ratio comparison of actual grades to estimated grades. This latter criterion will be discussed in

detail in the next paper (Froehlich). Other definitions include the use of the Accomplishment Quotient, the bright, (I.Q. over 130), children not getting all A's and B's, and average to bright children obtaining failing marks. Additional separation points for achievement are the Dean's list and probation lists (14, 49), scholarships and prizes awarded (92), or, in general, any other indication that a student is performing below or above his expected level.

Some investigators have made their comparison of high and low achievement without holding intelligence constant (52, 88), while others (31, 34) have systematically controlled all factors except class standing, which implies that only the extremes of the continuous achievement variable have been isolated and measured. Gough (36) presents a fairly good criticism of studies without adequate controls and points up the difficulties of comparative comment obtaining from these oversights. It becomes clear, then, that the proliferation of meanings, however precise they are intended to be, sometimes obfuscates the issues and impedes communication.

In terms of the criterion, some degree of deviation from the predicted grade point average (GPA) has been used most often as the indication of over- and underachievement. We might mention a few of the problems connected with this choice. The GPA is determined, of course, by the ratio of grade points earned to the number of academic hours taken. However, a student could increase his average from 72 to 79 and still obtain a "C" for the course grade, thereby not affecting his grade-point-average a whit. It would seem, therefore, that this criterion is too gross to pick up subtle, yet significant changes in performance. Moreover, it is well known that a host of other factors which may not be measurable by the usual battery of ability, interest and personality tests can affect grades sometimes by as much as two letters. Examples of these would be personality clashes with the instructor, grading quirks of the teacher (some may give only "C's" regardless of performance), tardiness or unexcused absences, or, perhaps, slouching in one's desk. On the positive side, higher grades than those actually earned may be given to the prospective "major" or, in the case of the co-ed,for having good-looking legs. The pleasant borderline student may get the "C" while his more surly classmate gets the "D." In some cases, their test scores in class might even be identical. All these factors contribute to variation in the criterion and gloss over possible student gains, yet are not likely to be measured by currently available

predictors. In spite of these deficiencies, however, grades are probably our best single criterion for classroom performance.

Other indices of academic achievement reported in the literature undoubtedly stem from grades but are used in different metrics. These include teachers' ratings, actual scholarship grants, prizes and peer ratings. In some cases, later professional or business success has been used as the ultimate criterion and, in these cases, academic success is reduced in importance as a yard-stick. It can probably be said that students who do well in school tend to do well beyond formal education. However, the skills demanded outside the classroom do not always fit into the criteria predicted by the usual academic aptitude battery.

From the data obtained on the discrepancy scores between expected and actual performance, the general consensus appears to suggest that non-intellectual factors may account for the unexplained variance in GPA not predicted from ability measures. Operating on this assumption, a wide range of personality, interest and attitude tests have been administered to achievers and non-achievers with the dual purpose of isolating those characteristics which differentiate the two groups and constructing scales which might predict achievement level after ability is removed from consideration. Whether attitude toward study or actual study habits is the superior index remains a moot question. Fakability is not the least of the pitfalls facing the examiner who relies on questionnaires when measuring college students. Because of the difficulty in divorcing fact from fancy with these instruments, students filling out questionnaires should, perhaps, be trained in much the same manner as were the introspectionists of some decades past. Nevertheless, the advocates of study habit inventories (17, 18, 19) and attitudes toward study (15) remain steadfast in their conviction that studying is important in academic success.

The preponderance of research on non-intellectual variables in achievement studies looks to personality measurement in its various forms, be it in terms of traits or dynamics, as the better solution. Within this framework, considerable emphasis has been placed on the Minnesota Multiphasic Personality Inventory (MMPI) and some assessment of need-achievement given, usually, by either a variation of the Thematic Apperception Test (TAT) or the Edwards Personal Preference Schedule (EPPS). The Rorschach has been tried in some instances (49), but the somewhat variable connotation of the determinants with respect to more

measurable attributes of the individual vitiates much of its predictive power. The cross-validation of Rorschach symbol frequency with estimates of achievement or leadership has consistently yielded low relationships. While there is no intention of questioning the test's construct validity in this paper, its predictive validity remains suspect. We shall concern ourselves, therefore, with studies incorporating the MMPI, EPPS, and TAT in an attempt to find correlates of over- and underachievement.

With the MMPI, most studies seem to indicate that students classified as underachievers tend to score higher on the Ma, Pd and Mf scales of that instrument, although when scales are constructed which are heavily loaded with items from these sub-tests prediction with a cross-validation group is not really successful. There may be some indication that over- and underachievers are not really characterized by specific personality patterns insofar as those factors which may reduce the output of some students may, for another, contribute to to his success. Some of the confusion resulting from MMPI studies may be a function of the different methods used in treating the data. Thus, t-tests between scale means yield information of a different sort from an analysis of possible patterns of scores. Similarly, men and women are sometimes combined on the Mf scale even though high scores on this scale mean opposite things for the two sexes. This latter observance has given rise to the notion that any analysis of over- and underachievement must necessarily allow for sex differences and any scale constructed on a combined sample will distort the picture when used with either category alone.

Another source of difficulty arising from the use of these personality devices is the fact that the tests have been developed and verified on a clinical sample out of which may develop construct or concurrent validity data. It is quite probable that information of this kind does not readily lend itself to placement in a prediction battery. To the extent that the tests are homogeneous, they may be less effective in tapping the criterion which, as we have seen, is often gross and, perhaps, overly broad. Some factors which may lead to good clinical devices may work to the detriment of prediction.

One difficulty in research on achievement which may be shared with other types of psychological investigation is the acceptance of different measuring devices as being equivalent. In this connection it would be well to consider the need to achieve as opposed to actual achievement in the

study of over- and underachievement. As Blake and Mouton[1] point out, the need of the year is Achievement with the rest of the AAA contributed by Anxiety and Authoritarianism. It appears that the TAT was doing quite well in terms of researcher comfort until Edwards came along with his Personal Preference Schedule. While those working with the thematic approach probably still feel comfortable about achievement, some authors have seemingly equated the two tests on the basis that since both mention need achievement they should certainly measure the same thing. We find a certain amount of research in this area hoping for some type of rapprochement. This, for the most part, has not proven to be the case. Most intercorrelations between the TAT and EPPS have not been significantly different from zero. Sometimes other need achievement scales are introduced which are also uncorrelated with each other. This has led to another conclusion that both tests are measuring n-ach but at different levels. Thus, the EPPS may tap manifest achievement needs (high value achievement) while the TAT defines more subtle aspects of the achievement drive. This has resulted in some confusion about the nature of n-ach and its relationship to actual achievement.

In this connection, a study by Parris and Rethlingshafer (75) compared students at different achievement levels (GPA) in terms of the McClelland n-ach test and found no support for the hypothesis that, other things being equal, high and low achievers will reflect high and low achievement needs respectively. I am not sure that one could even expect this to hold any more than one might expect that the religious person can be defined by the amount of money put into the collection box. We might suggest that n-ach may be reflected in a certain attitude toward tasks which is defined by the individual rather than by an outside authority. It should also be mentioned that, for the underachiever, conflict with an outside authority—the teacher—who is often an over-achiever may compound the former's difficulties which may virtually guarantee his continuing as an underachiever. Investigators measuring achievement are defining it in terms of goals established by others. An individual may feel he is achieving in some area, although he may be considered a failure by the outside criterion.

Having discussed some problems associated with the definition of, criterion for and the prediction of achievement, we may briefly consider some of the inferences

---

[1] Blake, R.R., and Mouton, J.S., "Personality." In Annual Review of Psychology. P. R. Farnsworth and Q. McNemar (Eds.), 1959, 10, 203-232.

drawn from over- and underachievement research. The underachiever tends to appear somewhat hostile, hostile especially to authority figures, and especially to authority figures who are teachers. At the same time, he is considered to be nonconformist, although some research indicates that this characteristic typifies the overachiever. The underachiever seems to have been brought up in highly authoritarian homes with excessive discipline which may tend to produce the passive-aggressive type of personality which often characterizes the underachiever. The underachiever socializes sometimes to the point where his studies may seriously suffer. While overachievers are sometimes guilty of this, they, in turn, tend to be more serious-minded, introverted, self-conscious and self-confident. The achiever, in general, seems to reflect greater maturity in both his work and outside activities than the underachiever.

In conclusion, we might mention that within our system, there is a certain emphasis placed on grades which most students, as they develop academically, learn to accept. Indeed, it is not uncommon for some students to dedicate themselves to earning good grades even if it must be at the expense of learning. Yet, individuals of no mean ability continue to falter in academics. A bright boy could fail high-school and become quite competent in a variety of criminal ventures. As is well known, many prominent statesmen, artists and writers have had difficulty in coping with the problems of American history, mathematics or literature and composition. We might also suggest that loafing through life successfully has been a notable achievement in which some members of our society have taken considerable pride. If our object is to ultimately understand the drive for and acquisition of achievement, we may have to reorganize our thinking about criteria such that the assessment of achievement must be in terms of the individual's criteria rather than our own. And insofar as the criteria are individual, perhaps our future methodological emphasis should vacate the nomothetic camp in deference to an idiographic assault.

DR. DuBOIS: I should like to ask whether the studies reported appeared to you to be rather short-range, opportunistic undertakings, or whether there has been systematic planning in the area.

MR. HACKETT: It is quite possible that many studies end too soon and at a point where they might profitably begin. A few people, among them Gough (36) and Gebhart and Hoyt (34), have criticized the seeming lack of concern with

tying up one study with another in the same area. They have also attempted to present a systematic outlook for these problems. Blake and Mouton[2] have also pointed out various discrepancies in these studies and the accompanying confusion.

DR. HARRIS: The sampling in these studies of convenience presents a problem, too. Did the investigators in the studies you reviewed pay much attention to this question?

MR. HACKETT: Some individuals have felt that there are sampling difficulties attending the use of certain tests. Specifically, problems of sampling associated with the Taylor Manifest Anxiety Scale often prevented its use in this context. However, I do not think there has been much specific recognition of the problem.

DR. HARRIS: Another technical problem, especially in opportunistic research, is that there is no cross-validation to check out the sampling problem. I had a student at Wisconsin who mined a lot of MMPI data, trying to build a non-intellectual predictor for university achievement. We found a number of items that were uncorrelated with ACE but correlated with achievement. But, not to make fools of ourselves, we did take the second step of cross-validating the results on a second sample. The thing went as flat as a day old soufflé.

MR. HACKETT: Many authors are well aware of the need for cross-validation. Quite a few of them cross-validate their studies and, consequently, find negative results.

DR. MAYO: On that same point, I have the impression that this problem was involved in most of the studies you were reviewing. Toward the end of your paper, however, there seemed to be some well-structured descriptions of the behavior of under- and overachievers. Do these studies lack cross-validation?

MR. HACKETT: Certain aspects of personality or interests seem to differentiate under- and overachievers rather well in some studies, but the characteristics do not seem to be consistent from one study to another. Some investigators indicate that the underachiever's grades suffer because of his high sociability, while others state that the over-

---

[2] Op. cit.

achiever is the gregarious type and the underachiever is introverted. Certainly, the usually high Ma score of underachievers on the MMPI tends to support the first interpretation. Perhaps one of the difficulties is the equating of indices for meaning. Does the Ma scale really furnish evidence of sociability, or does it indicate "expressiveness," which may be different? It could, of course, suggest the level of control over impulses and thus be more dissimilar. If one should use the Guilford-Zimmermann for some index of sociability and equate it with the Ma scale on the MMPI, one may be introducing more problems.

DR. CRONBACH: One of the difficulties is that the investigations almost invariably have been in a linear model and that, no matter what the statistics, we are trying to improve a multiple correlation. And yet all psychological and social class theory would argue that there is a passive type of underachievement which comes from not giving a damn about doing well and just getting by. But the aggressive, maladjusted type of underachiever is combined with the passive type by means of the usual criteria so that the basic way of organizing the data could not possibly make sense. One cannot deal with this problem unless people are sorted into homogeneous groups on some of these variables before appraising their achievements under appropriate circumstances. I do not know of any study that approaches the problem in a statistical way that is different from impressionistic. Is that your impression from the literature?

MR. HACKETT: Certainly, in terms of the personality variable, if you light a fire under two people, one may run while the other may turn around and put the fire out. This goes back to one of the primary difficulties in this type of research, viz., the lack of a consistent definition of what an over- or underachiever is. Moreover, once the type has been defined according to some criterion, there may still be equivocation on the terms used in characterizing him.

DR. JONES: While there is ample indication of a significant correlation between predicted and actual success in the classroom, it was suggested by many of the studies reviewed that the unknown variance in the criterion can be accounted for in terms of non-intellectual measures. But perhaps a lot of the variance which is presumably determined by non-intellectual predictors is actually the result of the error introduced, as Cronbach suggests, by not classifying

these subjects in homogeneous groups. Has anybody gone far enough to attempt to examine the variance carefully enough to know which part can be accounted for?

MR. HACKETT: The variance associated with each individual predictor can be isolated and the variance predicted by the combined battery can be determined. The addition of some scale, developed on criterion groups, to the test battery may increase the multiple correlation and account for some unique variance in the criterion. The highest multiple R reported in the literature is somewhere in the neighborhood of .65 or .70. However, it would be difficult to say whether the criterion had been de-contaminated to the point where one could state with assurance that the variance accounted for by the predictors was pure and could readily be broken down and identified by predictor.

DR. BRYAN: Concerning the criterion, are you talking about school grades per se or deviations between estimated school grades and actual school grades?

MR. HACKETT: In terms of the multiple correlation, the criterion would be school grades or grade point average. The deviation score would be used to differentiate, that is, define, the achiever and non-achiever.

DR. CRONBACH: The correlations involving achievement prediction have gone up about ten points in the last ten years. After the war, college marks were predictable to the extent of .45 or .55 and it was unusual to get outside that range. Now, particularly as the result of using something better than the ACE and using relatively longer batteries, it is commonplace for people to report correlations of .70. While there are obviously factors of unreliability in the grade average, some grade average criteria might be reliable insofar as a person is consistently poor from one course to another. That is, over several courses one might find reliability. Students may have illness in their families which could account for grade fluctuations in single courses, but this is not something for the psychologist to predict. There is an inherent limit of predictability. It seems to be more and more questionable how much we ought to hope to push up these correlations.

John French has attempted to predict differential performances in specific courses using a battery of factor tests, which have severe limitations because of brevity. He used the proper cross-validation design which,

after correction, gave results which predict grades in specific courses to the extent of about .80. The predictors were the more obviously pertinent measures such as aptitude and ability. Now, if careful psychometric methods get the correlations up to that level, it is very doubtful whether we can, by correlational methods, go any further.

MR. HACKETT: We might also mention that a lot of studies have related interest patterns with specific course grades. Yet, an individual with sufficient talent may enroll in a course and, not liking it, still do well. At the same time, if he were less endowed, he could like the course and still do poorly, so that the resulting correlation between interests and grades is attenuated. Perhaps the only way to really tap this variable would be to study these people who continue within a particular field as opposed to individual courses.

DR. CRONBACH: French was using the A2A of the Progressive Education Association, which is really a rather overt measure of, "Do you like history courses?", etc. This shares some of the faults of tests like the Brown-Holtzmann Inventory, which carries a considerable loading of past history. If you have been achieving badly, you will answer the questions in certain ways. To ask, "Do you like to take history tests?" or "Do you usually have trouble in school?" will yield a phony sort of prediction in some unassessible degree.

DR. TRAVERS: How does French get correlations as high as .80?

DR. CRONBACH: He goes beyond predicting single courses, and is taking people who have had two years of courses within a single field.

MR. HACKETT: To amplify Dr. Cronbach's remarks about identifying homogeneous groups as opposed to a heterogeneous classification of underachievers, we sometimes find that a student with a high quantitative score may do predictably well in mathematics or physics, yet, in a course in statistics, lose all conception of time, space, and content. Perhaps the ability tested becomes virtually irrelevant when placed in the context of different skills demanded by the situation.

DR. BRYAN: What is your attitude toward continuing efforts in this over- underachievement domain?

MR. HACKETT: There should be a greater exploration of what is meant by achievement, which inevitably has to be tied up with needs. The theoretical foundations are still hazy. As a consequence, much of the research seems to have little theoretical grounding, and to have emerged from off-the-cuff hypotheses. I suspect that if more basic work were done first, we might find more consistent results in the literature. Basic work must include attention to definition, both of predictors and criteria. The Edwards Personal Preference Schedule need-achievement scale has consistently been found to be unrelated to need-achievement on the TAT, yet investigators continue to use either as a measure of need-achievement, perhaps because the test authors say that is what their scale measures. Similarly, grades are looked upon as achievement indices, yet some students seek their satisfactions in life in areas other than grades. If money is not important to someone, can we say that he is unsuccessful if he has no money?

DR. BRYAN: You would feel, then, that this is a potentially productive area to work in, and you would not, as a result of your survey, wash your hands of the whole affair and look for more orderly domains?

MR. HACKETT: I would wash my hands of certain approaches because they are unproductive. For example, one study consisted in a year-long evaluation of a social worker's visits to the home of an underachiever to talk with his mother. After the year, the grades of the student were observed to see if there had been any increase. The conclusion was that the visits by the social worker did not help the underachiever's output. I suspect that all that could be said in that circumstance would be that a particular social worker was unsuccessful in altering the grades of a particular student. It would be nice if we would eliminate the grossness of our criteria and define our concepts more rigidly. Prediction by personality measures, which were designed in such a way that they could measure only a small fraction of any variance associated with success, seems futile. A more deliberate analysis of what it is that we are trying to measure may yield less activity but more action.

LT. FROEHLICH: In comparing under- and overachievers on MMPI scales and other personality inventories, using analysis of variance or t-tests, it is assumed that underachievers have a unique and abnormal pattern of traits and overachievers have a different and normal pattern. But, while significant differences may be found on one or two

scales, the patterns taken as a whole do not indicate any abnormality. By using a clinical scale on a "normal" population, they have assumed that underachievers must be low on some scales and overachievers high. But, in line with what Dr. Cronbach was saying, an individual could be low on a particular scale and be an underachiever, while another could be low on the same scale and be an achiever. I think research, whether using t-tests or analysis of variance, confined to demonstrating mean differences on MMPI scales would not be really fruitful.

Some investigators at the State University of Iowa are using a pattern analytical approach, within an analysis of variance framework, to isolate over- and underachievers. Instead of categorizing an individual as an over- or underachiever on the basis of an index determined by a regression equation, they hope to identify the over- and underachiever within certain types.

MR. BERKSHIRE: Enough work has been done with the MMPI that we are in a position to enunciate a Multiphasic law. If an item analysis of Multiphasic items is performed against any criterion, 5% of the items will be significant at the .05 level the first time the analysis is made.

DR. THORNDIKE: I am somewh favorably disposed toward the MMPI. I have a doctoral student who is cross-validating the items from the original MMPI scale on a new sample of Minnesota cases. What surprises me is that the subtle items hold up as well as the more obvious ones.

DR. GOFFARD: It seems to me that, because of a lot of traditional biases, research has been directed more toward what we should think about underachievers than what we should do about them. It is also clear that we should not think about underachievers but about an underachiever, for, if we want to do anything about it, we have to do it clinically. What should we do about underachievers as a purely practical problem?

MR. HACKETT: In some cases, the basal metabolic rate has been used as a diagnostic guide and, for a few, medical treatment may increase the energy level sufficiently so that they can at least open a book now and then. In the majority of cases, attitudinal problems interfere with the student's work and the treatment of choice is

counseling. We must recognize, however, that with the regression equation there are always as many overachievers as underachievers, regardless of the level of ability we are studying. Hence, as a practical matter, only those students who are failing or who are operating two letter grades below their potential are likely to come to the attention of responsible authorities. Certainly, action would not be taken on an individual merely on the evidence that he falls within an underachiever category.

DR. BRYAN: Thank you, Mr. Hackett. The next speaker is LTJG Herbert Froehlich of the Naval Technical Training Command, who will speak on the "PAQ as a Measure of Under- and Overachievement."

LT. FROEHLICH: For more than four decades psychologists and educators have attempted to define the non-intellectual characteristics of overachievement and underachievement and have thought that these characteristics would account for some of the variance in criterion scores unaccounted for by ability measures. This has implied interest in two major questions:

1. What are some of the non-intellectual characteristics of over- and underachievement?

2. How much do non-intellectual factors add to the prediction of academic achievement?

These still remain two of the more important questions. Overlooked, however, throughout these many years of dealing with overachievement/underachievement, although present in some form in practically every study, was an achievement index unhampered by the problems which surround the use of non-intellectual variables. This paper traces the development of this index and illustrates its use.

Studies which sought to compare the means of over- and underachievers on such scales as the MMPI or the Edwards PPS (Griffiths, 1945; Altus, 1948; Owens and Johnson, 1949; Morgan, 1952; Gebhart and Hoyt, 1958), or those which sought to construct an overachiever/underachiever scale through item analysis of such tests (Darley, 1937; Owens and Johnson, 1949; Gough, 1949; Frick, 1955), have used many different methods in isolating their criterion groups. The method of isolating criterion groups, however, was always a means to an end rather than an end in itself, and so the simple index with which this paper is concerned was overlooked. Needless to say, the correlations between non-intellectual variables and

achievement found in past studies were not very great, and occasionally when they were, they would not hold up under cross-validation.

In working with over- and underachievement we have overcome many of the problems which have plagued others. First off, we did away with non-intellectual variables, item analyzed or not. Second, we had rather large sample sizes when considered in relation to studies in this area which have used about 100 in each criterion group. Last, we did not eliminate the middle group of "par" achievers. We used a simple but neglected approach which previously might have been used to isolate criterion groups, but now became a variable in itself. That is, we devised our index, Preparatory School Achievement Quotient, (PAQ), as the ratio of an actual grade to that grade predicted by an ability test. This gave us an over-achievement/underachievement index for each student, one that measured that part of performance not associated with aptitude. We now had a non-intellectual predictor of achievement.

## Background

Before presenting data from our electronics and mechanics schools demonstrating the use of the PAQ score, we might briefly put the index in an historical perspective. The idea is not new. In 1939 DuBois suggested an achievement index based on the ratio of grade point average to ACE score, both measures reduced to standard score form. Guilford (1941) realized that DuBois' statistic assumed that grade point average and ACE were perfectly correlated, rather than only to the extent of .442, as reported, and suggested a ratio based on the obtained score to the expected score expressed in terms of a regression equation. This ratio yielded a mean of unity and a correlation of zero with the denominator (DuBois, 1947).

DuBois (1947) has indicated that achievement ratios should be freed of the factor of intelligence. The ratio we have been discussing, the PAQ, is free from the factor of intelligence as measured by the Navy Basic Test Battery, and is one whose variance is contributed by other factors. Mayo (1956) suggested that some of this variance may be contributed by student effort. In his study he found that the correlation of peer ratings on effort with the achievement ratio was .40. However, when the effect of peer ratings on intelligence, which correlated .66 with peer ratings on effort, was partialed out, the remaining correlation between peer ratings on effort and the achievement ratio was .16, which was significant at the .05 level. Mayo concluded

that even with halo strongly operative in the case of peer ratings on effort, a significant portion of the variance in the objective measure of effort, the achievement ratio, that is predicted by peer ratings on effort is not accounted for by peer ratings on intelligence. It should also be noted that in this study the achievement ratio correlated .10 with an objective measure of intelligence. A correlation closer to zero was expected.

As part of a larger project Mayo (1957), using the achievement index, showed that individuals who have psychological problems tended toward underachievement. This tendency was significant at the .05 level in a 1953 group of students and significant beyond the .01 level in a 1954 group.

Two of our main considerations in designing the measure of overachievement/underachievement have been to reduce the relationship between ability and achievement to a single term, and to eliminate the extraneous variable, intelligence. There are, of course, other ways of eliminating the variance contributed by intelligence. DuBois (1947) has proven that a difference score is similar to the ratio and so will also eliminate the factor of intelligence. In investigation of study habits Garcia and Whigham (1958) have confirmed this to the extent of showing that a difference score, gotten by subtracting a predicted grade point average from the actual grade point average, correlated close to zero with the predicted average and the ability measures.

In discussing the elimination of extraneous variance DuBois, Teel, and Petersen (1954) showed that if the variance which is considered extraneous is eliminated from two variables, the correlation of the residuals would be the partial correlation. This would be the same as the correlation of two ratios of the kind we have been discussing.

A procedure for partialing has been suggested by DuBois (1957). This method was taken up by Mayo and Manning (1961) in a study in which five variables which purported to measure motivation were evaluated in terms of their relationship to overachievement/underachievement. After all the variables were intercorrelated the variance associated with three aptitude variables was partialed out of each of the remaining variables and the resulting residuals intercorrelated. The removal of the aptitude variance converted an early measure of school grades and two later measures in a more advanced school to overachievement/underachievement measures. Although the underlying principle is similar to that described in this paper, it is strictly correlational and in practice does not provide an achievement index for each student.

## The Achievement Ratio

The achievement ratio about which we have been talking is put to use in the mechanics and electronics fundamentals schools and the more specialized "A" schools. In the fundamental mechanics school, which is only four weeks in length, the achievement ratio is based on the actual final average to the predicted final average. The index is then used in the "A" schools. In the longer fundamental electronics school (19 weeks) the ratio is the actual phase A grade to the phase A grade predicted by the Navy Basic Test Battery (General Classification Test, Arithmetic Test, Mechanics Test), abbreviated BTB. A regression equation with BTB predicting phase A was derived using 1450 trainees:

$$\text{Phase A Grade}_1 = 14.8280 + .3586\ (\text{GCT} + \text{ARI} + \text{MECH}).$$

The correlation between these two variables was .47. Another regression equation, for use in the "A" schools, was based on the graduates of the fundamentals school, excluding 133 students who dropped. The BTB and phase A grades correlated .42 and the regression equation was:

$$\text{Phase A Grade}_2 = 30.6213 + .2769\ (\text{GCT} + \text{ARI} + \text{MECH}).$$

The predicted score was divided into the actual score made by the man yielding an achievement score designated as PAQ, Preparatory School Achievement Quotient. The $PAQ_1$ may be used in the fundamentals school and the $PAQ_2$ in the "A" schools.

Table 1 shows the intercorrelation among the fundamentals school variables, including the $PAQ_1$ score. From this table it may be seen that the PAQ score predicted the

Table 1

Intercorrelations Among Fundamentals School Variables, Means and Standard Deviations for 1450 Students

| | BTB | $PAQ_1$ | Phase A | F Av | Mean | SD |
|---|---|---|---|---|---|---|
| BTB | -- | .01 | .47 | .50 | 178.98 | 11.53 |
| $PAQ_1$ | | -- | .88 | .68 | 1.00 | .10 |
| Phase A | | | -- | .84 | 79.00 | 8.85 |
| F Av | | | | -- | 77.51 | 8.00 |

final average somewhat better than the BTB. When PAQ and BTB are combined in a multiple to predict the fundamentals final average, a multiple correlation of .84 is gotten. This was to be expected since PAQ is constructed to correlate approximately zero (.01 here) with the Basic Test Battery and so contributes uniquely to the final average. The regression equation for this prediction is:

$$\text{Fundamentals Final Average} = .34\ (\text{BTB}) + 54.00\ \text{PAQ}_1 - 37.34.$$

Two-thirds of the obtained scores will be within 4.40 points of the predicted final scores.

Although the regression equation is sufficient for use in the fundamentals school, it was found desirable to develop a table from which predicted grades could be read by the school's personnel without difficulty. The table is such that all one has to do is enter it with the BTB score, read across the top, and his PAQ score, read along the side, and read off the student's predicted final average in the body of the table. A student advisor could use such a table to estimate how much better a student with a certain BTB score might be if he could be motivated to put our more effort. Thus a student with a PAQ score of .90 (an underachiever) could be encouraged so that he might be at least a "par" achiever, if not an overachiever.

Once the individual has graduated the fundamentals school the PAQ developed on the fundamentals course ($PAQ_2$) is used in the "A" schools. Again prediction is better using PAQ and BTB than either alone (reference may be made to Table 2 which illustrates data for one of the schools). It is of value to note that the correlation between $PAQ_1$ and $PAQ_2$ ranged from .986 to .991 for six schools. In light of such high correlations only the first PAQ is necessary for practical purposes.

Other Characteristics

With a PAQ score based on the actual fundamentals final average, rather than phase A, it is possible to get even higher multiple R's in the "A" schools. Table 2 shows this for one of the "A" schools. Table 2 shows this for one of the "A" schools. Three PAQs were developed in the ATN "A" school using the fundamentals phase A grade, the fundamentals final comprehensive grade, and the fundamentals final average. The following may be concluded from this table:

1. The BTB correlated practically zero with the three PAQs as expected.

2. The final average PAQ correlated higher with the "A" school final average (.69) than the phase A PAQ (.42). This suggests the use of the former in the "A" schools and either the latter or $PAQ_1$ in the fundamentals school.

3. The multiple correlation using BTB and the Phase A PAQ to predict the final "A" school average was .54; using the final comprehensive PAQ and BTB it was .65; and using the final average PAQ and BTB, .76.

4. The correlations among the PAQs were not as high as might be expected if overachievers were to remain overachievers throughout the training.

5. The correlations among the three PAQs were the same as the correlations among the grades on which they were based with ability partialed out (found from Table 2 by DuBois' method [1957] ).

Table 2

Intercorrelations Among Three PAQs and Grades for 523 "A" School Students.

Avionics Fundamentals

| | Phase A | Phase III | F Comp | F Av | $PAQ_2$ | $PAQ_{fc}$ | $PAQ_{fa}$ | "A" F AV | Mean | SD |
|---|---|---|---|---|---|---|---|---|---|---|
| BTB | .42 | .40 | .40 | .48 | .01 | .03 | .04 | .34 | 180.36 | 11.23 |
| Phase A | -- | .59 | .60 | .79 | .91 | .48 | .69 | .53 | 80.26 | 7.54 |
| Phase III | | -- | .70 | .84 | .46 | .60 | .78 | .71 | 79.18 | 7.27 |
| F Comp | | | -- | .83 | .48 | .93 | .75 | .60 | 77.89 | 8.56 |
| F Av | | | | -- | .65 | .71 | .90 | .75 | 79.04 | 6.46 |
| $PAQ_2$ | | | | | -- | .52 | .74 | .42 | 1.00 | .09 |
| $PAQ_{fc}$ | | | | | | -- | .79 | .52 | 1.00 | .10 |
| $PAQ_{fa}$ | | | | | | | -- | .69 | 1.00 | .07 |
| "A" F Av | | | | | | | | -- | 76.31 | 6.63 |

Data from six mechanic "A" schools were used to divide students into overachievers and underachievers. The division was at the mean of PAQ. A comparison of final average means for these two groups showed differences significant well beyond the .002 level. Although for three of the schools differences between the groups on BTB were noted at the .05 level, these differences were not consistent or systematically in favor of the overachiever group or the underachiever group. It might also be mentioned that prediction of the "A" school final average was no better in one group than the other.

## Discussion

This paper demonstrated that a successful achievement index could be developed unhampered by hypotheses as to what personality or interest factors distinguish overachievers from underachievers. In spite of its simplicity, this index has been used infrequently. However, this index, or any other, assumes that people are either overachievers or underachievers -- that overachievers are the precise opposites of underachievers. An index makes no allowance for the differential effect of, for example, personality on achievement. Thus, some variables might act as an aid to achievement in the case of one individual, or type of individual, and as a hindrance in the case of another. A pattern analytical approach (Haggard, 1958; McQuitty, 1959) might be recommended in which it is possible to predict the different types of overachievers and underachievers that researchers have hypothesized exist (Gebhart and Hoyt, 1958; Krug, 1959).

## Summary

The origin and past use of an overachievement/underachievement index has been briefly outlined, and its formulation as a ratio presented. Data supported the idea that an achievement index could be developed free of the ability factor and which, when added to an ability measure, improved the prediction of school grades.

The intercorrelation of ratios of actual to predicted grades was seen to be the same as the correlation among grades with ability partialed out. Although overachievers may differ from underachievers in their psychological makeup, it was supposed that a pattern analytical approach is required to isolate the different types of over- and underachievers which may be hypothesized to exist. Overachievers differ significantly from underachievers in their course grades. PAQ was shown to be practical but nevertheless a gross approach to isolating and studying over- and underachievers.

Altus, W.D. A college achiever and non-achiever scale for the Minnesota Multiphasic Personality Inventory. J. Appl. Psychol., 1948, 32, 385-397.

Darley, J. G. Scholastic achievement and measured maladjustment. J. Appl. Psychol., 1937, 21, 597-606.

DuBois, P. H. Achievement ratios of college students. J. Educ. Psychol., 1939, 30, 699-702.

DuBois, P. H. On the statistics of ratios (Abstract). Amer. Psychologist, 1947.

DuBois, P. H. Multivariate Correlational Analysis. New York: Harper, 1957.

DuBois, P. H., Teel, K. S., and Petersen, R. L. On the validity of proficiency tests. Educ. Psychol. Measmt., 1954, 14, 605-616.

Frick, J. W. Improving the prediction of academic achievement by use of the MMPI. J. Appl. Psychol., 1955, 39, 49-52.

Garcia, D., and Whigham, N. Validity of SSHA administered before and after college experience. Educ. Psychol. Measmt., 1958, 18, 845-851.

Gebhart, G. and Hoyt, D. Personality needs of under- and overachieving freshmen. J. Appl. Psychol., 1958, 42, 125-128.

Gough, H. G. Factors relating to the academic achievement of high school students. J. Educ. Psychol., 1949, 40, 65-78.

Griffiths, G. B. The relationship between scholastic achievement and personality adjustment of men college students. J. Appl. Psychol., 1945, 29, 360-367.

Guilford, J. P. A note on DuBois' method of deriving achievement ratios for students. J. Educ. Psychol., 1941, 32, 220-222.

Haggard, E. A. Intraclass correlation and the analysis of variance. New York: Dryden, 1958.

Mayo, G. D. Peer ratings and halo. Educ. Psychol. Measmt., 1956, 16, 317-323.

Mayo, G. D. Differentiating characteristics of a group of students having psychological problems. J. Educ. Psychol., 1957, 48, 359-370

Mayo, G. D. and Manning, W. H. Motivation Measurement. Educ. Psychol. Measmt., 1961, 21, 73-83.

McQuitty, L. L. Differential validity in some pattern analytical methods. In B. N. Bass and I. A. Berg, (Eds.), Objective approaches to personality assessment. Princeton: D. Van Nostrand, 1959.

Morgan, H. H. A psychometric comparison of achieving and non-achieving college students of high ability. J. Consult. Psychol., 1952, 16, 292-298.

Owens, W. A. and Johnson, W. C. Some measured personality traits of collegiate underachievers. J. Educ. Psychol., 1949, 40, 41-46.

DR. BRYAN: Thank you, Lt. Froehlich. Are there any comments or questions?

DR. THORNDIKE: You have taken your total prediction of a level of achievement and have broken it up into two pieces, and have made your prediction in two parts rather than one. But have you gotten any better prediction than you would get from the whole?

LT. FROEHLICH: These two pieces, of course, account for all the variance in the predictor, and, when combined, account for the same variance in the criterion as the undivided predictor. However, the use of the PAQ as a measure of over- and underachievement is valuable to counselors. When they see a man's phase grade, Basic Test Battern scores and PAQ, they can take action. If they see a man who is an underachiever, they can pull him out of class and send him to night school. Perhaps a high ability underachiever can be given encouragement and guidance, or maybe dropped; a low ability overachiever, although not doing very well, may be given a chance to repeat part of the course.

DR. THORNDIKE: The question is whether the counselors would do any better by using the discrepancy between the Basic Battery and phase grade as the basis for putting a student in a special school than by randomly assigning poor students for special instruction. That is, you are assuming that the discrepancy score is a reliably established and meaningful basis for taking action with respect to the individual.

LT. FROEHLICH: Yes, Dr. Mayo has shown that PAQ is related to effort and, certainly, the schools are always asking students to apply themselves more. I suggest that if we assume that the low PAQ of an underachiever is related to a lack of effort perhaps, by motivating him in some experimental learning situation we can, theoretically, make him a par achiever or even an overachiever. However, many of the studies I reviewed indicated little change in grades following counseling.

DR. THORNDIKE: Are you suggesting that picking up these boys who are underachievers and sending them to night school is, perhaps, a futile enterprise?

LT. FROEHLICH: Looking at this from the viewpoint of the counselor or educational consultant, or anyone on the selection board, when we select individuals

out at random, how do we know which one is going to benefit from night school? We would like something more substantial so we do not have to guess.

However, we can see how effective the PAQ actually is in isolating those people who are truly underachievers. And, by applying appropriate motivation to them, we can follow them up. The procedure of randomly selecting students from a course simply on the basis of low grades and putting them in night school may be too clinical. At the moment, no such random selection is taking place. All students who fail a phase examination are sent to night school. We hope that PAQ might identify students who would or would not benefit from this special instruction.

DR. THORNDIKE: You can still follow up on it, however.

LT. FROEHLICH: We do not have grades or an index of the extent to which these people are underachievers. I think PAQ fills a need in that respect. There is an index called "Q" which is used in some schools. This is determined by subtracting 30 from twice the student's course average and dividing the difference by the sum of the GCT and ARI, two of the Basic Tests which average 120 for students in that school. Thus,

$$Q = \frac{2\bar{X} - 30}{GCT + ARI}$$

Assuming the course average is 75, a par achiever would have a "Q" index of 1.0, since

$$Q = \frac{150 - 30}{120}$$

But, here again, the numerator and denominator are assumed to correlate perfectly. The instructors in one of our advanced schools feel it has been highly successful in picking out the underachievers for special instruction.

DR. THORNDIKE: Successful in the respect that it makes the counselor feel more comfortable?

LT. FROEHLICH: Well, it does that too, of course. The primary objective is to get as many students through the course as possible. We do not have the means of selecting people with the highest ability to send to these

schools. Now, if 10,000 students come through the program, a quick index is needed to see which will be over- or underachievers. When these indices are calculated on these thousands of students, a roster is sent over to the school based on the PAQ. The counselors can then see immediately whether a particular man is working up to his potential.

When dealing with thousands, it is impossible to counsel each individual on a person-to-person basis. PAQ makes the work easier. The counselor must advise whether to drop the student from school, or send him to night school for two weeks. This, of course, involves a considerable expenditure of money. Since developing PAQ in 1957, however, we have found that the people in the schools consider it to be quite successful. We can, of course, determine its value in the improvement of grades, satisfaction with the Navy, etc., by continually following up the material provided by the schools. In the meantime, though, it is a rough index as to who will be an over- or underachiever and what his predicted final grade will be.

DR. MAYO: I think there may be some impression that more is being claimed for this index than is actually the case. We know that we can take a sample of students' performance and predict a sample of performance similar to the work that they may be doing at a subsequent time. We know that we can predict better with PAQ and aptitude tests than we can with aptitude tests alone. We can divide the initial performance into two parts: the part which is perfectly correlated with aptitude measures, and the part which is not. We can express that in the same metric in which we express their aptitude scores and make it readily available for administrative purposes. Contained in the part which is unrelated to aptitude are all other sources of variance in the criterion, some of which we may be interested in measuring. And we have demonstrated that this part correlates with peer-ratings on effort as well as with subsequent under- and overachievement as the man comes in the school.

DR. JONES: Assuming the utility of the index, do you ever try to use cutting scores of any sort? In the guide you furnish your counselors, it seems to me you would want to take as much advantage of chance as possible.

DR. MAYO: You would concern yourself only with those who deviate a good bit from their expected performance level.

DR. HARRIS: One point that bothers me is that if we separate variables into components, there are a number of ways of doing it. Most of the analyses of the data are dependent upon what kind of scheme is used for setting up differences. I would suggest there are some quite different analyses which have a rather elegant characteristic in that the analyses are invariant over what method is used in pulling something out. This is very important, because, if the results are a function of the way in which the separation is made, then practically nothing is learned from them, since someone else could do it differently and get different results. It would be better to look for those conditions under which invariance over such measures is obtained.

DR. BRYAN: The next paper will be presented by Mr. Roger Berkshire of the Naval School of Aviation Medicine. The paper is entitled, "Under-Measurement of Intelligence."

MR. BERKSHIRE: The proper way to begin a research report is with a review of the literature on the problem. Typically, the scientist sets out to convince you that he first surveyed all of the pertinent studies, that from the results of these studies he drew inferences as to the probable nature and organization of reality, that he incorporated these inferences into formal hypotheses, and that he then designed his experiment to test these hypotheses.

Regrettably the activities I am about to describe were not like that at all. Our hypothesis developed, rather inadvertently, during a coffee-break. And it seemed to us so immediately charming, so loaded with social significance that we avoided looking into the literature on the matter for fear of finding out that the whole idea was nonsense.

This particular coffee-break conversation was about the reported inferiority in the average measured intelligence of Negroes — and the probable effects of this in integrated schools — when the following notions suddenly developed. First, we were generally agreed that the score a person makes on an intelligence test is affected both by his heredity and by the environment (or culture) to which he has been exposed. Second, it seemed plausible to us that heredity might be contributing something like "capacity," or "ability," or "talent," while the culture contributed skills and knowledges specific to the intelligence test content. If these things are so, then it is possible to think of two people who have equal intelligence test scores; but one of these people

comes from an environment which contributed substantially toward his score, while the other comes from a culturally impoverished environment (for our purposes here, an "impoverished culture" is defined as one which fails to provide its members with the correct answers to intelligence test items). Since their scores are equal, one can reason that the person from the poorer environment must have had the better heredity — for he did as well on our intelligence test by virtue of heredity alone as did the other with the assistance of his superior environment. Further, if you are going to permit us to define what is inherited as "capacity," or "ability," or "talent," then he has the more "capacity" of the two. And if this "capacity" can be expected to operate for him in the future as it apparently has in the past, then he might be expected to learn more in new situations than his more culturally favored test-mate, despite their equal intelligence test scores.

So one way of stating our hypothesis might be, "If you match on intelligence test score people who come from widely different cultural (or perhaps socio-economic) backgrounds, those from the poorer background will do better in new learning situations." Or, in terms more suitable to this conference, "If intelligence test scores are used to predict performance on learning tasks, people from poor environments will appear to overachieve, while those from good environments will appear to underachieve." Of course, if this works out, then over- and underachievement become identified not as behavioral characteristics of individuals but as artifacts of the cultural aspect of our intelligence tests.

Now the enormous social significance of this hypothesis at this particular moment in history should be clear. Nationally, it means that Negroes and other out-groups will do better in school than their tests scores predict. Locally, it means that the participants in this conference might as well pack up and go home, because the hypothesis says that the kinds of data that are the subject of this conference do not indicate overachievement at all; they just indicate under-measurement. All that remains is for me to produce a little supporting evidence. Unfortunately, it is at this point that the hypothesis tends to lose a little of its glamour. The portion of intelligence test scores which can be affected by cultural differences must be relatively small; perhaps, (from studies of identical twins raised in different homes), something on the order of ten IQ points. Thus the effects of the hypothesized preponderance of genetic influence that we might find in one of a pair of matched scores would also be relatively small. Further, the supporting evidence must be drawn from

learning measures that have at best forty per cent and at worst about five per cent of their variance in common with intelligence test scores. This created pretty long odds against our finding our hypothesized difference. So we seemed to have a hypothesis which was charming, but of uncertain virtue. Therefore we adopted the standard operating procedure for situations in which extreme charm is associated with dubious virtue — we conducted an exploratory study.

Our experimental population consisted of men with two or more years of college who were entering naval air training. To get an estimate of their socio-economic background we administered a questionnaire concerning conditions in their home at the time they were in high school. The items for this questionnaire were taken from the American Home Scale and the Simm's Socio-Economic Rating, modified to suit our age group. We included items on books and periodicals in the home, education of parents, parents' memberships in such things as PTA or professional societies, parents' leisure activities — such as contract bridge, golf, tennis, ping-pong, music lessons, etc. For our intelligence measure we administered Form I of the Naval Aviation Qualification test. Despite its name, the AQT is just a pretty good verbal-numerical intelligence test. Now we needed some learning situations. We didn't know whether we expected the hypothesis to work best in short learning tests or in sustained performance, as in a classroom course. So we tried both. We made up three very short machine-scorable tests called Symbol Learning, Syllable Learning, and Serial Learning. We also administered the DuBois-Bunch Learning Test. And for sustained learning, we had the grades our subjects made later in pre-flight school. Also, a peer-rating of leadership potential was available to us, so we included it.

Table I shows the intercorrelations of these measures for a sample of 396 naval air trainees. If you will look at the second row, you will see that the intercorrelations of the culture score with the learning measures are mainly negative, as they should be if our hypothesis is to be supported. Of course, you can also see that they are too small to be very exciting. The largest negative correlation is with the aerodynamics grade. Because this correlation is negative, holding the culture score constant results in increasing the correlation of AQT and aerodynamics from .390 to .402. As my teen-agers would say, "Big deal!" However, one of the things we have learned at Pensacola, in recent years, is that correlation coefficients of low significance sometime conceal highly significant information. Or, said another way, in problems of primary and secondary selection, the pay-off relationships

usually involve only a relatively few cases at the extremes of the distributions. Besides, the data from which we make our living seem to include an appalling number of non-normal distributions and non-linear relationships. For instance, in addition to the positive skew that generally results from selection we have noted the following phenomena in past collections of data: men who received top ratings as fleet officers had high AQTs, but men who received low ratings had personality problems; men with very low AQTs had low probabilities of completing training, but so did men with very high AQTs; the same thing was true of men who received low and very high peer-ratings; very low pre-solo flight grades predicted failure in the fleet, but high grades had no relationship to degree of success. Correlation matrices, while apparently very satisfying esthetically to recent graduate students, are poor devices for discovering relationships like these.

So before we abandon our gaudy hypothesis, let us run our cases through the sorter. First, we will restate the hypothesis, but just for the extremes. We will say that if a boy comes from a very poor environment and has a very low intelligence score — this is what you expect. But if a boy from a very good environment has a very poor score — he must be really stupid. Conversely, if a boy comes from a very good home and has a very high intelligence score, this is also what you expect. But one who comes from a very poor home and makes the same score has overcome some considerable handicaps. So we sorted out the groups that fulfilled these conditions to see how they stood on our learning variables. First we took all cases with AQTs of 90 or above and 70 or below. We divided each of these groups into those with culture scores above 21 and those below 14. This gave us the numbers of cases shown in Table 2. Table 3 shows the mean scores of these extreme groups on the experimental variables. For the hypothesis to be supported the means in the low culture columns should be substantially larger than those in the high culture columns. The only variable for which this is true is not a learning variable at all — but the peer-rating that we put in only because it was free, not because it had anything to do with our hypothesis. This peer-rating is obtained in the eighth week of training from men who have been living and working together under rather stressful conditions. They are asked to pick the three best and the three worst men in their section (15 to 30 men) for battalion commander and to give their reasons for their choices. The pooled ratings normally correlate from .35 to .45 with subsequent failure, and the odds are about three or four to one that men rated in the bottom six per cent will not complete training. Note that this validity is for

training completion, not for leadership—although the latter is what the peers are presumed to be rating.

Figure 1 shows the correlation surface for this peer-rating variable. The mean peer-ratings of each of the segments are given on the surface and the percentage of high and low ratings in each segment are shown below. (A high rating is 60 or above, a low rating 40 or below.)

It looks as if men were perceived as having average and better leadership potential if they were of high intelligence and from poor environments and as having average and worse potential when they had low intelligence and came from good homes, but the magnitudes of the differences seem far too large to fit the small differences in "capacity" that our original hypothesis would lead us to expect.

So we went back and looked at the statements made by their peers about these classes of men, and found that the high-rated, high AQT, low culture men were said to be well organized, highly motivated, mentally aggressive, hard workers, enthusiastic, highly interested, and to have a good military attitude. About the low-rated, high AQT men it was said that they lack effort and determination, do not care about themselves or classmates, do not have motivation to fit their intelligence, are satisfied with less than their best, are unable to assume responsibility, etc. It is clear that the perceived differentiating factor among these high AQT-score men is mainly level of motivation — which is in turn negatively related to the reported cultural level of their homes.

When we examine the comments made about low-rated, low AQT men of high culture we find much the same thing, but in somewhat different words. "No spirit, no initiative, no desire to become an officer, does not want program, not enough go, whines, no self-confidence, sloppy, duty shirker, etc." Here again the differentiating factor appears to be primarily motivation.

Now if we accept these peer judgments as evidence of real differences in motivation, we are forced to rewrite our original hypothesis. It now becomes, "If you match on intelligence test score people who come from widely different cultural backgrounds, those from the worse background will tend to be better motivated and will frequently do better in new learning situations."

And all of a sudden we seem to be playing in the same ball park as the other teams present here. In fact, we may even

be playing the same game. But we have got an option on a shiny new bat called culture, which seems to us to be a necessary part of the game. We claim only an option on it because this whole thing needs to be validated by further studies, which we intend.

Table 1

Intercorrelation Matrix: AQT, Culture and Learning Variables
(N = 396)

| | AQT | Cult | Symbol | Syllable | Serial | DuBois-Bunch | Aero | Naval Orient | Nav | Eng | Peer Rating |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AQT | | .06 | .11 | .07 | .23 | .01 | .39 | .32 | .44 | .23 | .14 |
| Culture | | | -.06 | -.04 | .08 | -.01 | -.13 | .04 | -.11 | -.10 | -.09 |
| Symbol | | | | .27 | .20 | -.02 | .16 | .07 | .16 | .10 | .11 |
| Syllable | | | | | .23 | .06 | .12 | .12 | .04 | -.08 | .14 |
| Serial | | | | | | .06 | .13 | .01 | .17 | .00 | .02 |
| DuBois-Bunch | | | | | | | .01 | -.07 | .03 | -.01 | -.04 |
| Aerodynamics | | | | | | | | .32 | .58 | .47 | .21 |
| Naval Orientation | | | | | | | | | .36 | .23 | .11 |
| Navigation | | | | | | | | | | .41 | .28 |
| Engines | | | | | | | | | | | .10 |
| Peer Rating | | | | | | | | | | | |
| Mean | 80.2 | 17.5 | 18.5 | 13.4 | 19.3 | 62.9 | 45.1 | 48.3 | 48.4 | 49.2 | 51.0 |
| S.D. | 12.3 | 6.6 | 2.2 | 3.7 | 7.8 | 32.3 | 9.3 | 8.4 | 6.8 | 8.3 | 11.9 |

Table 2

Numbers of Cases with Extreme AQT and Culture Scores

| | Low Culture Score (Less than 14) | High Culture Score (More than 21) |
|---|---|---|
| Low AQT (70 or Less) | 31 | 22 |
| High AQT (90 or More) | 39 | 34 |

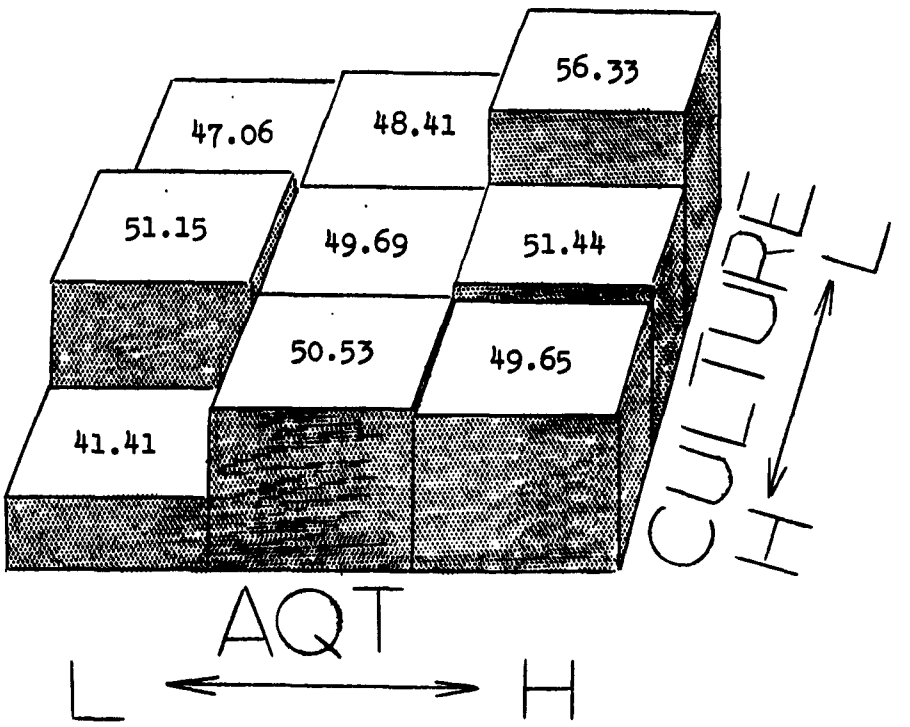
Table 3

Mean Learning Measure Scores of Extreme Culture Groups with High and Low AQTs

| | Low AQT | | High AQT | |
|---|---|---|---|---|
| | Low Culture | High Culture | Low Culture | High Culture |
| Symbol | 18.00 | 17.35 | 19.05 | 19.11 |
| Syllable | 11.97 | 11.91 | 13.95 | 13.29 |
| Serial | 14.89 | 16.48 | 21.69 | 21.97 |
| DuBois - Bunch | 72.50 | 58.26 | 62.31 | 67.06 |
| Aerodynamics | 39.75 | 40.80 | 51.68 | 49.71 |
| Naval Orientation | 44.76 | 43.73 | 50.84 | 52.91 |
| Navigation | 44.48 | 42.68 | 54.18 | 52.41 |
| Engines | 47.21 | 46.65 | 52.92 | 51.79 |
| Peer Rating | 47.06 | 41.41 | 56.33 | 49.65 |

Figure 1

Correlation Surface for Peer Rating Variable




Mean peer rating of each segment given on surface; percentages of high ratings (60 or above) and low ratings (40 or below) in each segment are shown below.

| % High | % Low | % High | % Low | % High | % Low |
|---|---|---|---|---|---|
| 3 | 16 | 13 | 30 | 25 | 3 |
| 9 | 9 | 12 | 11 | 28 | 14 |
| 9 | 40 | 13 | 13 | 20 | 14 |

DR. BRYAN: Thank you very much. Any comments?

DR. THORNDIKE: Is there a possibility that this relationship between motivation and culture is specific to the Navy's situation?

MR. BERKSHIRE: The relationship between motivation and culture may be specific to the Navy. A boy from a poor background may think that to be a Naval officer and a pilot is a good way to get ahead.

DR. THORNDIKE: Is this a real and significant objective for them?

MR. BERKSHIRE: I am sure this is a big part of it.

DR. THORNDIKE: Are the raters as a group more sympathetic to low culture than to high culture groups?

MR. BERKSHIRE: No, we have a very wide range of scores on the culture variable and the distribution is more or less normal.

DR. THORNDIKE: I have talked to students in classes involved in peer ratings and they have often told me the devices that the group had engaged in to jimmy the ratings. There may be a kind of sympathy with the man who is trying to get ahead, making a more friendly rating situation for the low socio-economic man than for the high socio-economic person.

MR. BERKSHIRE: We know of cases in O.C.S. where groups, in order to protect themselves, have by collusion agreed upon the ratings in advance so that everyone could come out even. We have no evidence of this happening in our group. Moreover, we have a very high attrition rate among people who receive very low peer ratings.

DR. JONES: You were quite careful to point out that you have a unique kind of failure group. In addition to those who are academic failures, there are those who are voluntary failures, who must have motivation to quit. I wonder if you would care to comment further on this aspect in flight training. There are not many places in which you ask a man if he wants to quit.

MR. BERKSHIRE: About 15% of those entering Naval aviation training decide to quit when they

first come into contact with an airplane. Furthermore, the identity of these people is known, to some extent, by means of peer-ratings taken earlier. What is even more predictable from peer ratings is who is likely to fail the flight training program. Some of the voluntary withdrawals are anxiety or fear cases, while others shorten their military obligation in order to return to school to prepare for a different career.

DR. LORDAHL: Can you assume that two years in college prior to entry into the Navy could help those in the low culture group catch up?

MR. BERKSHIRE: Our usual population is quite homogeneous, but this group is less so. While they do come in after two years of college, there are various reasons for their doing so. Some of them were not doing very well in college, others left because of other factors. Hence, we have a less homogeneous group here, with considerable spread. It is nothing like what is expected in the general high school population.

DR. CRONBACH: I would like to look at something you have not worried about much, back here in Table I. I think this represents a phenomenon that is going to need more thought. I do not know the reliability of the various criteria, but it is obvious that the prediction of grades from AQT is appreciably higher than the prediction of the learning measures. And this relates to the theoretical point that Woodrow made. This should not have occurred if we were truly measuring aptitude. We still look upon the aptitude measure as indicating something about one's equipment he can bring to bear on a new task. Yet, we see the differences between the courses and your learning measures. Something of a similar nature is implied in Fleischman's work on motor skills where we find a pretty good prediction early in training but poor performance at the complex level where, within the training, inconsistent individual differences develop.

Obviously, aptitude as we customarily measure it by sampling procedures does not tell much about what happens in a fairly homogeneous learning context. Now the only place where I know this has been demonstrated in a practical learning process is in the new physics curriculum. The Physical Science Study Committee has worked out a unique curriculum for high school physics. The Educational Testing Service has prepared a group of achievement tests, all of which are much the same type. This is a series of ten tests, each dealing with a different segment of the subject matter: light,

or atomic energy, mechanics, and so on. The achievement tests are given about every six weeks and correlated with the post measure. There is a clear decline in correlation, as you go through the course, of these achievement measures with the predictive scores provided by SAT. In a course which is essentially heterogeneous, that is, consisting of segments of material, the aptitude measure is a good predictor. If, on the other hand, the program is highly cumulative, then the prediction will be poorer and something that looks like specific factor variance will develop during the training program.

Now, I do not know whether in other physics courses, or in any other course, measures over time would result in similar divergence. I do not think a study has been done with repeated achievement tests. However, if I am right in my speculations, a sampling approach to measurement results in decent prediction because of the lack of the accumulative characteristic. This phenomenon has been found in two rather good measurement situations, and it ought to have a lot more attention.

DR. BRYAN: The discussant for this morning's session is Dr. Robert M. W. Travers, University of Utah.

DR. TRAVERS: This is not a field in which I have any degree of expertness, so I am something of an amateur looking in from the outside.

The thing that does impress me most is that we have had tremendous difficulty here in building a model on which we can develop a program of research. We still seem to be at the stage where we have to scout around and collect as many facts as we can in the hope that, somehow, some day, all these will fit into some kind of model. We are not fussy about whether the things we are dealing with are the consequences of this variable under- and overachievement or whether they are antecedents and causes of the condition.

I am tempted to try to build some kind of model. While listening to the discussion, I was wondering about what kind of intervening variable model could be generated in this area. I think the best candidate for an intervening variable would be the motivational variable, such as achievement

need, anxiety, or, perhaps, compulsiveness. I can see, though, why it is very difficult to build any model of this kind. I have had experiences in trying to measure achievement needs and I recently had my fingers burned. When you are measuring this kind of a variable, it is like trying to determine the volume of a liquid when the liquid is highly sensitive to temperature changes and you have no control over temperature. This is apparently why we cannot get any reproducibility of results when we deal with achievement needs.

I have, for example, some data on two forms of the TAT, trying to measure need-achievement. I have zero correlation on the two forms based on TAT type pictures. This also has zero correlation with French's measure of achievement needs. I suspect that my conditions are so variable and so overwhelming that, again, it is like measuring the volume of liquid which changes rapidly with temperature changes, when you have absolutely no control over temperature. This makes me feel that, until we can get better control over conditions of measurement of some of these variables, it is almost useless to try to talk about them as components in an intervening variable model of achievement.

I feel that some kind of conceptual model might be more satisfactory in this area. If you know what a person's goals are, then maybe you can begin to determine some of the conditions through which he will achieve those goals. For example, the counseling center at the University of Utah says that one outlet for overachievement is to marry a wife who will glorify one's career. Included here is the over-glamorized wife, or the wife who may be, intellectually, somewhat higher up on the scale and who may help her husband to achieve goals otherwise impossible, such as in writing. This is a way through which they are able to lift themselves up by their bootstraps, but you can know that only if you know something about their goals and the means by which they achieve them.

There are lots of related phenomena of this kind. Some day, someone will put them together, but obviously they cannot fit into the intervening model at the time.

I was very much interested in the last paper. I suspect that Mr. Berkshire underestimates the influences with which helpful background can affect I.Q. scores. He puts it at ten points. While the average difference is ten points, as you go to the extreme differences in the cultural environment of identical twins, you get up to 20 or 24 points, probably a sigma and a half, which is quite large. This fits well with data on

Negroes who came from the South, but were raised in Harlem schools. Over the years their I.Q.'s went up 10 or 12 points. I am interested in how these people manage to acquire motivation to be able to build their own environment and not just to accept the environment they have.

MR. SPIES: In speaking about variability, how much of it is inherent within the individual that is being measured, and how much is a function of the measuring instrument? If within the majority of individuals, this variability were great just by chance factors, then it would not be too worthwhile to keep refining the techniques for an overall measurement when more clinical approaches would be necessary. On the other hand, possibly there is more long range stability within individuals. The variability within individuals would be quite important to know.

DR. DuBOIS: I wanted to ask Roger about his learning measures. Are these measures of change, or are they total measures? What sort of measures do you have for "symbolic learning?" Also, how did you score the learning test that Marion Bunch and I made up some time ago?

MR. BERKSHIRE: We tried scoring the learning test in several ways, one of which I think you recommended. That was to subtract the median of the first two trials from the median of the last two and get the gain scores. As for the other three tests, they were too short and unreliable, and, therefore, unsatisfactory.

Conventional memory for symbols was measured by a test in which subjects are given six minutes to read and study a list of syllables and then respond on a machine-scorable answer sheet.

The slope of the learning curve was determined with the DuBois-Bunch test. The other short tests would not give us a curve. They yield only a single answer, a single score, since the initial performance is subtracted from the final score.

There was a little gain, however, since one of the troubles was that, on the second trial, too many hit the top on the test. I would appreciate suggestions as to better learning measures to be used in a repetition of this kind.

MR. HACKETT: We are currently developing a series of learning tests which may have some general use for studies of this nature.

MR. BERKSHIRE: One thing I plan to use is a paragraph learning test in which they will be given paragraphs of moderately complex technical materials, and perhaps fifteen minutes to study this, and, then, without having reference to the paragraph, will answer questions on it. This would be simulation of the usual school situation.

DR. CRONBACH: I am having some problem in determining what you are trying to measure. Is it what we call immediate memory, or can it be comprehension? In this last example, I do not see that we are really getting learning. If the test is immediate, what are we getting?

MR. BERKSHIRE: We have used this paragraph for a test of comprehension. Typically, we let the student have access to the paragraph. He goes back and checks the paragraph for his answer. This way we define it as comprehension. However, without recourse to the reading material, it approximates the typical school situation except for a smaller time lag between the time studied and the time of examination. But this still seems to be learning. It is the kind of learning I seem to have done a lot of.

DR. CRONBACH: What I am really asking is whether learning in the short time and learning on the long term basis are the same?

MR. BERKSHIRE: They may not be. When we do it again, we will still include school measures.

DR. BRYAN: In psychomotor tests, it seems fairly easy to keep track of the course of learning. With the kind of material that you are dealing with, it seems that you have not hit upon some scheme for doing it.

MR. BERKSHIRE: I still think the initial hypothesis is charming. I am not ready to give it up, but I am not convinced that we did it well.

DR. LORDAHL: I suggest the type of material used in the operation would make a difference. I would not expect a correlation between idle conditioning and intelligence or between serial learning and intelligence. You might only find the learning curve could be more rapid in the above cases. But, certainly in the manipulation of symbols and abstractions, I would expect a difference. If you start out with something relatively unique or meaningless to both groups, then the rate of improvement and the degree of manipulation would be

better for the high intelligence group than, in the context of your hypothesis, for the low-culture group.

DR. CRONBACH: It seems to me you might get a learning measure of an intellectual sort somewhat comparable to the score on the motor function. Finding how far the person had gotten after a period of time, the question would be how much generality that would have, how much relation the progress on the one task would have with an entirely different task.

DR. TRAVERS: I think there is quite a lot of data to show the rejection of Berkshire's hypothesis. We did work of predicting achievements on various cultural groups and found that the southern Negro from a deprived environment performed at the level expected in terms of his measured I.Q., but not in terms of the level expected with his corrected I.Q. Whatever depresses measures of general intelligence also depresses achievement in academic and other learning situations. And I can find no evidence to the contrary.

DR. THORNDIKE: In the Bronx High School of Science in New York City, they place the underachievers with the low socio-economic group and the achievers with the high socio-economic group. In this instance, the socio-economic variable was a predictor. You put it in with a positive rather than a negative weight in a very specific kind of context.

DR. TRAVERS: Some data on teachers indicate that, while students of education constitute a heterogeneous group, their grade achievement is considerable better than what is expected in terms of prediction. They are overachievers. They appear to be a group struggling upwards in the socio-economic scale.

DR. THORNDIKE: Are they also overachievers in mathematics and physics, or only in education?

DR. TRAVERS: They are overachievers in mathematics and physics also. This is a very general phenomenon. We do not have as wide a spread in the educational courses as we should have.

MR. HACKETT: Among the characteristics of underachievers is usually an avoidance of the mathematical and technical areas. Overachievers tend to gravitate towards these areas rather than underachievers. Certainly, the prospect of success in calculus may carry more prestige than that found in some social studies.

DR. CRONBACH: One of the things that has not gotten properly into the discussion, because most of the military is not concerned with it, is the sex variable. But your literature review, I suppose, was done mostly on mixed sex groups and makes interpretation difficult. Regarding the relation at certain age levels, am I right in saying the investigator used both sexes, generally, and did not separate them in the analysis?

MR. HACKETT: Some of the more recent investigations have separated them.

DR. THORNDIKE: In the lower grades girls are overachievers and boys are underachievers.

DR. TRAVERS: It may be more of an environment factor. I would expect that the environment is better for the girl. The elementary school is run by women, she is in a woman's society, and it puts the boys at a real disadvantage in developing in that kind of environment. They tend to reject so many things about it.

MR. HACKETT: And, according to the literature, the girls tend to achieve more in relationship to their ability in general than do the boys. Also, males tend to emerge as underachievers early; from grade one, in fact.

DR. HARRIS: I have not heard the fine word, "creativity," considered here this morning. To what extent are the kinds of measures we have dependent upon the point of view we have? There is some difference between responding to a vocabulary test and some kind of ideation which allows one to make lots of responses to a set of restricted topics. There are devices in measurement procedures which lend themselves to this type of research. In fact, this might be more profitable than going down the personality trail with the MMPI.

MR. HACKETT: One study that directs itself specifically to this topic was done by a girl with the charming name of Janet Eugenia Puccinelli Wallerscheim — which is creativity in its own right - who measured the relation between creativity and achievement in grade and high school. She found, using modifications of the Guilford tests, that the achievers consistently scored higher on all the creativity tests.

DR. HARRIS: This is an area in which, at least for certain age groups, you would better pay a lot of attention to the difference between samples of boys and samples

of girls. The overwhelming finding is that the picture of the inner relationships of these types of performances for boys and girls are quite different. They cannot be brought together. I suspect Guilford is finding this now.

DR. TRAVERS: Has anybody been able to predict anything with a creativity measure? While Guilford's work is valuable, I cannot find any predictions based on it.

DR. HARRIS: One study obtained a number of measures of creativity or versatility in terms of teacher experience with judgments of various types of projects that they had done in different subject matter areas. While some of the correlations were consistent with using creativity as a predictor, a factor analysis was performed in order to boil them down to a composite. The correlations were not too displeasing, being in the range of .20 to .40. Of course, there was some variation as a function of subject matter. You cannot be sure that projects in math are being judged on creativity in the same sense as those in the social studies. This is a big problem.

DR. BRYAN: If there are no other comments, we will adjourn for lunch.

## Afternoon Session, 14 April 1961

(The afternoon meeting was called to order at 1:30 p.m., with Dr. Marion E. Bunch, presiding.)

DR. BUNCH: The first paper on the agenda for this afternoon is by Dr. DuBois and is entitled, "Correlational Analysis in the Investigation of Under- and Over-achievement."

DR. DuBOIS: In my files I have a paper I wrote more than 20 years ago entitled, "On the Statistics of Ratios." It was written as the result of Guilford's criticism of a previous paper of mine and it was accepted by the Journal of Educational Psychology. About the time of the war I asked for the manuscript back to make a few changes, so as to generalize the approach a bit. Somehow the changes were never made. When I needed a paper for a meeting in 1948, I used it, but since then it has rested in the files.

I want to present a few ideas from that paper and then go on to what is a somewhat more generalized approach to the evaluation of achievement than the fairly simple ratio approach of 20 odd years ago.

When we are talking about factors in achievement, we have several models, of which the simplest involves a single predictor and linear relationship. It accounts for a good deal of the research which comes under the rubric of under- and overachievement.

There are several ways of combining two variables, X and Y, in such a way that the linear correlation between the combination and one of the variables (which I shall call the base variable) is zero. One of them is, of course, the familiar residual, $z_{x.y}$. That $z_{x.y}$ correlates .00 with $z_y$ is well known. There are two families of numerical solutions for developing these derived measures. "On the Statistics of Ratios" is concerned with the two linear combinations of two variables combined in such a fashion that the correlation with the base variable is zero. It includes a demonstration that statistically the four solutions, two of them ratios and two of them differences, are exactly the same. The correlations of the four solutions are all unity. Numerically, of course, they are different.

One solution involves the ratio $X'/Y$ developed in such a fashion that the correlation with Y will be zero. This

is accomplished by letting X' be a standard score with $M'_X = M_y$ and $\sigma'_X = \sigma_y / r_{xy}$.

A second ratio technique is to use $X/\tilde{X}$, in which $\tilde{X}$ is the score in X as predicted from Y by means of the regression equation in raw score form.

One of the two subtractive techniques is to use (X' - Y), in which X' again is the standard score with $M'_X = M_y$ and $\sigma'_X = \sigma_y / r_{xy}$. The other subtractive technique is the familiar residual, mentioned earlier, which may be used in the raw score form $(X - \tilde{X})$ or as a reduced z-score, $(z_x - \tilde{z}_x)$ or $z_{x.y}$.

Although determined on different numerical scales, all correlations of these derived variables are, of course, identical. In each case the linear correlation with the predictor variable is .00 and the correlation with the criterion is $\sqrt{1 - r^2_{xy}}$. In correlational analysis the residual is the most convenient, since individual values need not be found and all of its correlations can be precisely inferred from the complete correlation matrix of the original variables.

It seems to me that the rationale of the ratio and of related techniques is the development of a score which will correlate zero with the predictor, and still contain all the unpredicted variance of the criterion. On the other hand, much of the research on under- and overachievement has used the model of a single predictor, thereafter looking for subsequent predictors which will related to the unpredicted variance.

There is a variant in under- and overachievement studies in which there is a search for distinguishing characteristics between extreme underachievers and overachievers on some variable other than the one on which the group has been separated. The notion of using extreme groups, it seems to me, is not nearly as satisfactory as a systematic exploration of the various factors that can be identified in achievement.

The second model is one that has not been used very much. It might be called the single predictor model with non-linear relationship. The only example of such a derived measure I have been able to think of has been the Wechsler deviation I.Q. in which Wechsler is dealing with the non-linear facet of the relationship between scores on intelligence tests and age. At each age level he uses a mean of 100 and a sigma of approximately 15. By standardizing at each successive age level, he has taken out an effect of age in a non-linear fashion.

If we find such non-linear relationships between achievement measures and predictive measures, and if we are working with only one predictor at a time, this would seem to be an acceptable mode.

The generalization of the ratio technique is quite simple. Such generalization yields the third model, that of multiple predictors and linear relationships. From the criterion variable, any number of predictor variables are partialed out, yielding a residual which is unrelated linearly to any of the variables which we have partialed out, either singly or in combination.

The fourth model involves multiple predictors with non-linear relationships with achievement. In addition to first powers of the predictors we might use squares and third powers, and any other functions that seemed desirable. Computation might become complicated, but we certainly can investigate non-linear relationships in a multiple fashion if we wish to do so.

I am going to speak mostly about the third model because in our present status of knowledge it is one of the simplest with which to work in the analysis of achievement, Here we work with any number of predictors but consider only linear relationships. Of course, if non-linear relationships appear, they can be exploited.

In using the concept of under- and over-achievement, the implication is always that after one or more of the predictors have been used, there are one or more other predictors in reserve. If we have used all the conceivable predictive information that we have in our variables, then what is left is the error. And if we have no more predictors in reserve, I do not think that the term under- or overachievement really applies. The residual, $z_{0.12...n}$, is the unpredicted part of the achievement after n predictors have been used and there is no further way of identifying any part of it.

When this model is used there are some questions that are worthwhile answering. One is the unique predictive value of each of the predictive measures. Now, the unique contributions, of course, will not add up to the total prediction at all, but it seems to me that when we talk about a creativity test, or a personality test, or some other observed measure, that we would like to know whether all of the predictiveness of that measure can be accounted for by other predictors, or whether it has something unique. That is rather easily found by a part correlation.

The part correlation (sometimes called the semi-partial correlation) is an indication of the relationship between an unmodified variable on the one hand and a residual on the other. If Variable 0 is the criterion and Variables 1...n are the predictors, then the proportion of the criterion variance predicted uniquely by any variable is the square of the part correlation between the criterion and that predictor resiualized with respect to the other predictors. Thus the unique contribution of Variable 1 to $R^2_{0(12...n)}$ is $r^2_{0(1.23...n)}$, which may be found by dividing the appropriate partial covariance, $C^2_{01.23...n}$, by the variance of the residual, $V_{1.23...n}$. This procedure takes advantage of the fact that in z-score form a part covariance and a partial covariance are numerically identical.

The part and partial correlations, however, are not identical. The part r is a more logical analytic device in this case, since there is no reason to partial anything out of the criterion.

For a three-predictor problem one way of finding the 3 second order part r's is as follows:

| Unique Contribution of Variable | |
|---|---|
| 1 | $r^2_{0(1.23)} = R^2_{0(123)} - R^2_{0(23)}$ |
| 2 | $r^2_{0(2.13)} = R^2_{0(123)} - R^2_{0(13)}$ |
| 3 | $r^2_{0(3.12)} = R^2_{0(123)} - R^2_{0(12)}$ |

By analogy part r's of any order may be found. The can also be found from partial covariances and partial variances obtained during matrix reduction.

If there is no suppressor, the sum of the squares of the part r's will be less than multiple $R^2$. If there is a suppressor in the system the sum of the squares of the part r's will be greater than $R^2$. Nevertheless, the part r shows promise of usefulness in breaking up the criterion variance into components.

If we are interested in residual gain as our criterion, the correlations are still simple to work out. The fact that the criterion itself is a residual variable in no way changes the general analytic approach.

Within the frame of reference of a set of predictor variables, the total variance of a criterion can be broken up into defined portions, namely, what is uniquely predictable from each single source, what is predictable from two sources, and what is predictable from three sources, and so on. This type of analysis may be useful in trying to understand the underlying factors in a psychological phenomenon such as achievement. In any matrix our analysis is certainly relative. We are not going to get out of a matrix anything we do not put in it.

I have considered an alternate possibility in analyzing a criterion in terms of n predictive variables. This would be a method of analysis about halfway between regression analysis and factor analysis. If we were to have communalities so that each variable were completely predictable from all the other variables, then all multiple correlations would become unity. Then it would be possible to see configurations of meaning quite clearly.

DR. BUNCH: The next paper is by Dr. Robert L. Thorndike on "Methodological Issues in Relation to the Definition and Appraisal of Underachievement."

DR. THORNDIKE: My responsibility today is to talk about some of the methodological problems relating to over- and underachievement. I got into this enterprise when I undertook to provide a do-it-yourself guide for would-be educational researchers on this topic for the Cooperative Research Project of the United States Office of Education. This office has apparently been swamped with ill-designed projects attempting to deal with underachievement. My responsibility to them is to provide a guide that can be used to avoid some of the more common design inadequacies.

Unfortunately, the pressure of events during the past six months has prevented me from moving as fast or as far as I would have liked. Consequently, I anticipate that much of what I have to say will already have been covered by previous speakers, and much of the rest of it will be repeated by others who come after me. In a gathering such as this, I expect that we will see eye to eye with respect to many of the technical problems involved in dealing with this somewhat messy topic.

Basically, the problems of over- and underachievement can be viewed as just a special case of the general problem of multivariate prediction. We are simply trying to determine what accounts for individual differences in achievement.

Essentially, we are attempting to account for the variance in a criterion measure. I conceive of this variance as arising from four sources as follows: (1) error or measurement, (2) criterion heterogeneity, (3) variance determined by substantially unmodifiable aspects of the person or environment, (4) variance determined by relatively modifiable features in the individual's world. Let me expand briefly on each of these.

The notion of error of measurement is certainly a thoroughly familiar one to this group. Though different experimental operations result in somewhat different definitions of just what shall be treated as error, we mean roughly that variance in a set of scores that is uncorrelated with a second experimentally independent, but completely comparable set of scores. In all of educational research, but particularly in our thinking about this problem of over- and underachievement, error of measurement is an unmitigated nuisance. Error of measurement in the criterion variable means that we deal not with a true criterion, but with a fallible estimate of it. It means that we deal with the appearance of high or low achievement, sometimes without the reality. Since errors of measurement occur in the predictor as well as in the criterion variable, their combined effect can account for substantial individual discrepancies between performance on the two measures. That is, much of what appears to be over- and underachievement may represent nothing more than the combination of errors of measurement.

The reason that errors of measurement become particularly distressing to us as we try to study the problem of over- and underachievement is that when we make a selection of cases on the basis of high or low performance on a single test score, or especially on the basis of discrepancy between two test scores, we will get a loading of errors of measurement in one direction, and consequently systematic regression effects on any re-test. That is, if we pick a sample because it is low on a given testing with an achievement test, we may expect the members of the sample to regress upwards toward the group mean on a re-testing, even though we merely wave a magic wand over them between the two testings. Or if we pick a sample whose achievement test scores are relatively low and aptitude test scores relatively high, we may expect the achievement test scores to rise and the aptitude test scores to drop when both are repeated. Errors of measurement are bad enough when they are random, but in the typical design of the study of over- or underachievement the errors of measurement are not random but systematically biased, and they are likely to lead us to biased conclusions.

The second source of variance in criterion measures is what I have chosen to call criterion heterogeneity. Criterion heterogeneity differs from random error or measurement in that the variance in this second category can be systematically associated with some other fact known about individuals or groups of individuals. One fairly clear illustration of criterion heterogeneity is that which results from combining college grade-point averages for individuals in different schools within a college or university. Thus, for example, one might combine students in engineering and students in agriculture. I think we would all agree that a specific grade-point average would then represent quite a different level of academic attainment in the one case than in the other. Criterion heterogeneity could also be illustrated by the tendency to grade girls more leniently than boys, by differences in grading standards between a regular program and an honors program, by differences in standards between institutions, or by any other differences that are systematically associated with some fact known about the individual.

Criterion heterogeneity thus introduces systematic bias in the criterion score. When this occurs, the criterion-predictor discrepancy will be related to any factors that are related to this bias. Thus, if in some university, country boys tend to major in agriculture whereas city boys tend to major in engineering, a study of a joint population of agriculture and engineering students would be likely to show rural residence to be associated with overachievement and city residence to be associated with underachievement.

Whenever identifiable sub-groups exist with respect to the criterion score, and whenever one suspects that there may be heterogeneity in the meaning of the criterion for these different sub-groups, some systematic analysis should be made of this matter. The appropriate analysis would appear to be an analysis of covariance between the predictor and the criterion, carried out over the series of sub-groups. This analysis would make it possible to determine whether a common regression applied to all the sub-groups. If it were found that one did not, then it would be necessary either to adjust the criterion scores to make allowance for the heterogeneity in the criterion measure or to carry out separate analyses within each relatively homogeneous criterion group. A number of the researches in the field of under-achievement have recognized this problem and tried to deal with it, usually by restricting the group in some way, but a number of others clearly have not.

The third category of variance in criterion scores is variance that can be predicted from factors known about

the individual or his environment that are relatively unmodifiable during the educational process with which we are concerned. One such factor might be initial level of performance in the achievement area in which we are interested. That is, insofar as final status on the achievement measure is predicted by initial level of achievement, individual differences in training or experience during the experimental period do not appear to have any significant differential effect. Other unmodifiable factors would be such things as initial level of performance in various measures of aptitude, the sex of the individual, or the socio-economic level of the home from which he comes. Factors such as these are likely to provide the core of our multivariate prediction of the final criterion score. Typically, only one or two such variables have been used to define the expected achievement, and over- or underachievement has been expressed as the discrepancy from the achievement that would be predicted on the basis of this single initial score. The approach in terms of a single predictor seems inadequate. One phase of research on the underachievement problem demands that we explore all the facts that can be known about the individual in advance and that give promise of predicting his final achievement. This exploration would permit us to set up a team of predictors that would account for all or nearly all of the variance that can be accounted for in advance of the learning experience.

Insofar as the variance of the criterion scores can be accounted for by errors of measurement, criterion heterogeneity and permanent factors of the sort aforementioned, the problem of over- and underachievement largely disappears as a genuine problem. The situation then becomes one in which all of the predictable criterion variance is predicted by things that we are unable to change and in which, consequently, no margin exists for such factors as changes in individual's motivation, changes in the individual's learning skills, or the impact of particular types of learning experiences. However, if some variance remains that has not been accounted for by my first three categories, then we do have a genuine educational problem and once we have appraised the importance of such factors we can undertake to modify them in order to see what we can do to modify the educational outcomes. However, it is necessary for us to have a thoroughly realistic and complete appraisal of the variance accounted for by the first three components if we are to make a realistic approach to the issue of understanding individual differences in achievement and trying to overcome the deficiencies of achievement that appear in some individuals. If, as I suspect is often the case, only a very small fraction of the total criterion variance falls in this

fourth category, we are likely to be dealing in very fragile and insensitive experimental designs. Statistically significant results are likely to be hard to come by and to require large groups, and genuine experimental differences are likely to be small compared to statistical artifacts and biases.

As I size up the situation, there appear to be three possible strategies for investigation of the problem of over- and underachievement. I might call these respectively (1) the strategy of concurrent prediction, (2) the strategy of of genuine prediction over time, (3) the strategy of experimental intervention and manipulation. I would like to discuss each of these in turn and briefly.

In the strategy of concurrent prediction, we deal with a battery of variables all obtained at essentially the same time, in which we have designated one as the criterion variable and attempt to predict it from the others. This is essentially the strategy involved when we give an aptitude test and an achievement test and deal with the discrepancies between the two scores. The strategy is essentially the same whether we deal with a continuous distribution of discrepancy scores, or with two extreme groups that we label overachievers and underachievers. It is the one on which a very large part of the over- and underachievement research is based, because the fact of underachievement is very often inferred from this discrepancy between concurrent measures.

The strategy of concurrent prediction appears to me to involve one in a basic dilemma. This is the dilemma of differentiating between what is legitimately a predictor and what is really part of the criterion. For example, the typical scholastic aptitude test is likely to contain a test of vocabulary. So does the typical test of reading achievement. If we use the scholastic aptitude test as a predictor of the reading achievement test, the correlation between the two of them will arise in part at least because they include the same material. That is, the correlation between the two is only in part a matter of meaningful prediction and is in part a matter of the contamination of one measure by what is inherent in the other.

This type of contamination is most obvious when the two measures contain identical types of test materials. The contamination may be just as real when the two measures tap the same psychological functions, even though the materials are not identical. At least, it is hard to define where contamination stops and legitimate prediction begins. That is,

if cultural deprivation influences achievement on the one hand and similarly influences the aptitude measure on the other hand, a spurious relationship between the two has been induced by this outside variable, and no independent appraisal of cultural deprivation as a factor making for underachievement is really possible.

On the other hand, if one limits oneself to predictors that clearly do not have any overlap with or contamination by factors influencing the criterion variable, one tends to rule out most of the predictors to which one would ordinarily turn in order to get a genuine prediction of that criterion variable. In this case, the prediction will be very poor, and there will be much unpredicted variance in the criterion. This unpredicted variance will be a result of the pressure to avoid including in the predictor variables anything that could be conceived of as contaminated by the same factors that influenced the criterion.

This dilemma leads me to conclude that the strategy of concurrent prediction is a rather futile one as far as the problem of under- and overachievement is concerned. It seems to me that the results will inevitably be ambiguous, and that no really conclusive studies will be possible.

The second type of strategy is that involving genuine prediction over time. In dealing with genuine prediction over time, clearly we should be interested in the learning that has taken place during this period of time. That is, the meaningful prediction would involve not just the final status of the individual but rather the change of status of the individual from the beginning to the end of the prediction period. This means that for a clear-cut and meaningful experiment we should always have a measure of initial status in the domain represented by the criterion as well as a final score. The only instance in which this requirement could possibly be waived would be one in which we could assume that initial level of achievement would be uniform, which would normally mean that it was uniformly zero or so nearly zero as to make no difference. In any other situation, one of the most essential members of our team of predictors would be a measure of initial status, and we would be interested in other predictors primarily as adding to the effectiveness of the prediction that can be provided by a measure of initial status alone.

Evidence of the importance of initial status as a predictor in our pattern of predicting over time is seen in the fact that in any one of the standard school skills such as level of reading comprehension the correlation between an

initial and a final achievement measure is often very nearly as high as the reliability of the two measures permits. In the limit, when the correlation between initial and final status reaches the ceiling permitted by the reliabilities of the two measures, then there is nothing that can be added to an initial measure of achievement (excepting to extend it to a length when it becomes perfectly reliable) in order to obtain the best possible prediction of final achievement. It is important to determine that individuals have differed sufficiently widely in their rates of gain from initial to final test to permit some other variable to function as a predictor of this rate of gain. Only if there are reliably measured individual differences in gain does it make sense to look for other factors in order to relate them to this amount of gain.

There are <u>likely</u> to be reliable individual differences in gain in proportion as (1) a long time has elapsed between the initial and the final measure, (2) the achievement was a novel one so that all members of the group started at a near-zero level on the relevant skill, and/or (3) individuals differed widely in the effectiveness of the training they received. These conditions define the circumstances under which studies of over- and underachievement are likely to be fruitful. Unfortunately, they are not met in many of the studies carried out in public schools and colleges.

Two types of factors can be differentiated as possible predictors of gain. On the one hand there are factors that existed in the individual at the time of the initial test. These are comparable to the factors that I referred to in my third category as I was discussing factors in the variance of a set of criterion scores. Insofar as they are factors that characterized the individual prior to the learning period they are not subject to manipulation, and they serve merely to help us to understand how it is that some individuals gain more than others.

A second type of factor is a factor occurring or introduced <u>during</u> the learning period. Thus, if some individuals have special remedial instruction, while others do not, the presence or absence of remedial instruction can be treated as a variable and related to the final level of achievement on the criterion score. By the same token, if special attempts are made to motivate some individuals, or to provide alleviation of unfavorable home circumstances for them, this experimental variable can also be related to the final score. Variables occurring during the learning period may be variables that occur spontaneously, as well as ones that we introduce systematically.

Thus, the number of days absence due to illness might be a spontaneously occurring variable. Any of the variables that either occur or are introduced during the learning period can be correlated with final achievement, either as continuous variables or as categorical ones, and a correlational analysis can be carried out to determine whether consideration of these variables makes possible a more accurate prediction of our criterion score.

The discussion of variables that might be systematically manipulated during the learning period leads naturally into the third experimental strategy. This is the strategy of experimental control or experimental intervention; in other words, the strategy of introducing some special type of treatment. Though this kind of strategy can be handled within the correlational framework that I have just been discussing, it fits more naturally into the pattern of analysis of variance and study of the differences between discrete groups.

When one is operating with the strategy of experimental intervention, all of the usual demands of experimental method hold. Perhaps most crucial is the requirement for appropriate control groups. In practice, the critical problem is to get really equivalent control groups. In a great deal of the research in which some type of ameliorative treatment is applied in an attempt to overcome what is considered to be underachievement, the practical concerns of trying to help people have the inside track, and the development of experimental and control groups tends to take place after the fact rather than as part of the initial design of the enterprise. One group may be individuals who were included in a special program in which they were given special remedial help in reading, for example, and the other group may be made up of individuals who are now considered to have been equivalent to the experimental group in their reading disability but who for some reason were not in the special program. In such a case as this there is always a question as to whether the two groups did not differ in a number of subtle ways. Even if they are matched retrospectively with respect to external variables, there are likely to have been subtle statistical artifacts in the matching. I wrote some twenty years ago on the statistical difficulties of using matched groups, when the matched groups are chosen from what is not essentially the same population. In this after-the-fact type of matching one can rarely guarantee that complete equivalence is achieved. The only strategy for this type of experiment that is really completely satisfying is a strategy in which the total population of candidates for special treatment is found in advance, assignment to special treatment groups is then

made on a completely random basis, and some type of placebo treatment is provided for those who do not have what is considered to be the experimental treatment. Only under these circumstances does one have a real sense of conviction that the experimental treatment made a genuine difference.

With respect to the three experimental strategies that I have suggested, my tendency is to reject the first, that is, the strategy of concurrent prediction, as hopelessly ambiguous. My tendency is to be highly skeptical of the third, that is, the strategy of experimental intervention, unless the treatment groups are thoroughly defined in advance and assignment is on a truly random basis. My tendency is to prefer the second, that is, the strategy of prediction over time. In the second strategy I would insist upon a measure of initial achievement, and would then be interested in studying the additional factors that relate to final achievement.

In my discussion, special groups of underachievers have tended to be swallowed up in a continuium of achievement. The problem of underachievement is seen as continuous with the general problem of predicting achievement. It seems to be largely matters of practical concern rather than those of research strategy that set underachievement up as a unique and distinctive problem. Though it is possible that the prediction problem changes qualitatively as one moves down the scale of achievement and away from the regression line of achievement on aptitude, this is not necessarily the case. The strategies that I propose assume that the prediction problem is essentially one throughout its whole range.

DR. BUNCH: Having heard both papers we will turn to Dr. Cronbach for a discussion of these.

DR. CRONBACH: I found myself listening with considerable interest to these papers and only wish that I had about two weeks to think about what was said before trying to discuss them. My thoughts went off in several dozen directions. I am surprised that we have had so many papers today on the subject of under- and overachievement, since, obviously, it is a false problem. Both speakers this afternoon have agreed on the point that it is only a problem of how well one can predict an outcome. Perhaps it would be valuable to reflect, for a moment, on how we got into the bind. We have had a large amount of research on what is, essentially, a distorted form of the multivariate prediction problem. We can trace this back to the naive enthusiasm of the 1930's when intelligence tests were known to be measuring capacity, and were regarded as extremely

reliable; consequently, anything that could not be measured by such tests did not matter. Students either were or were not working up to capacity; hence, over- or underachievement.

The minute we take the position that has been suggested by these papers we are looking at the under- or over-achievement of the psychologist, rather than that of the pupil. There is some limit to our capacity to predict which is essentially inherent in the accuracy of the criterion and the standardization of the conditions under which data are gathered. Beyond this it is purely a question of how close we have come, with our investigative methods, to telling what will happen to people, to writing their biographies in advance. So, the separation of groups by the so-called capacity measures represents a faith in them we no longer hold.

The problem is also distorted by the view that we are dealing with fixed capacity. If there is just one capacity, then it does not matter when it is measured. When you admit that your behavioral samples are individual variables, some representing past learning and some representing performance in which the subject has had very little experience, the underachievement problem has to be defined more subtly.

Thorndike talked about some of the considerations, but he put them in an engineering context. If you want to predict who is going to be an underachiever in the school, and if you use aptitude as the predictor, there must be some cognizance of the extent to which any aptitude measure involves past achievement. Whenever we try to explain achievement, we are in a sad situation because the achievement we are trying to explain today is partly explained by some achieving that took place in previous years.

Because of the cumulative nature of experience there is no point at which underachievement begins, unless you want to count birth. Most people who have formulated research in this area have been working on better selection techniques for the schools. They have been operating naively under the assumption that past achievement is not relevant to present achievement. Thorndike says approximately what I would say here.

I would hesitate, however, to emphasize the purely statistical side of the problem as much as these two papers do. If our interest in the underachiever is to keep him from coming in, then we do want to combine tests with the ideal weight for eliminating the poor performer. But that is merely

the problem of predicting low achievement, rather than the problem of predicting underachievement.

DR. THORNDIKE: You would then be eliminating the low achievers. I would make a distinction between underachievement and low achievement.

DR. CRONBACH: I quite agree with you. If our interest is purely in doing a technical personnel management job then we have a statistical question and find the best predictors. Insofar as we are much more interested in understanding achievement, then I think the straightforward statistical approach is not necessarily the most satisfactory. I will make a variety of comments that will illustrate this.

I will first comment on what I think has been an historical bias in approaching this problem, present in the two papers we have just heard, that is the idea that we are using one fixed treatment. The job in education has always been to expose people to treatment in order to make them different from what they were. The educators, somehow, had the idea that they were selecting people for an educational program that was fixed and standardized. Hence, it would be meaningful to do research that would give them a regression formula for predicting success.

Now, there are two things wrong with this view. The first is that our educational program is invariably one of choosing between two treatments. Only in very rare situations are you going to reject people; you are going to educate them all. But the question is how? Rather than just looking at the single multiple correlation with one criterion, the approach that ought to concern us is the study of a separate regression surface for each treatment. This would permit us to compare alternative methods of instruction and assign the individual to whichever is most appropriate.

Thorndike talked about experimental techniques. I think there is a difference in point of view reflected in two or three places. For example, he said one group is given the experimental treatment, while the other group receives the placebo treatment. That is not quite ideal. You should consider the alternative educational methods that seem to be promising, and apply one to one group and one to another. I would agree on the advisability of using randomly selected groups, but the idea of treating one group and giving the other group the placebo treatment puts the emphasis in the wrong place. I think we should select treatments that fit the people and people who fit the treatment.

Something similar is involved in Thorndike's statement about how much the permanent qualities of the individual account for the criterion performance and how much is left subject to experimental variation as, for instance, the increase in achievement scores resulting from visits by a social worker. I would be far more concerned to look for the interaction between the experimental method and the permanent qualities of the individual. I think it is reasonable to suppose that some people will take well to teaching machines and some will not.

It seems to me we get into some difficult problems when we begin by working in a purely statistical frame of reference and say x will be the best predictor, y the next best, etc. I am sure many of you have seen Cureton's remarks in which he points out that in using correlated predictors, the weights assigned to particular predictors in the multiple will be largely or, at least, significantly affected by sampling variations. Obviously, if there are two correlated predictors, one will get a higher weight the first time around, but, with another sample the weight could be conceivably negative.

Carrying this further, I think that we get into enormous confusion when we start adding a great number of predictors and then analyse the predictors separately. Pat Sears was interested in predicting achievement in elementary schools as a function of several things, such as: class differences, pupils' personalities, various aptitudes, and past achievements. There are more variables than subjects, yet it is obvious that we cannot run comprehensive studies with thousands of subjects. With such a large number of variables, she can get bewildering differences between groups if she merely works out, for instance, the part correlations or beta weights and uses these as a solution.

It seems to me we have to maintain a fairly firm control in the experimental situation in order to draw conclusions that can be put into some sort of theory. My hunch as to how to control so that group differences are not unduly influenced by sampling fluctuations, is to reduce the variables as much as possible by imposing a factor analytical structure on the predictors. Thus, instead of starting with all the predictors, we probably should do some sort of factor analysis of the predictors to reduce the number before we start looking for their weights.

Having been warned by Dr. Harris that we should not try to separate variables, I know that what I am about to say is not the right answer. Nevertheless, what we have been doing in this situation is to use the square root or diagonal method of factor analysis whereby we group the variables in an attempt to pick out a general intellectual factor or verbal factor. This factor is defined by the average of two tests. The variance of that factor is then extracted from the predictor matrix and any small difference that might exist between the two tests would probably not show up in the predictor actually used.

This was also done with a sociometric device in which there were peer ratings of eight types. Rather than throw the eight predictors into the regression formula, it made better sense to pick out one thing that seemed to be significant, for example popularity, and define that by the composite of three choices, such as whom they would like to work with, or play with, or the like. The composite, as a predictor, simply replaces the three original values. Then, if we find that performance has a relation with this variable we will be able to interpret it. Under these circumstances, the regression surfaces will not be bewildering. When the engineer takes bewildering regression weights, puts them together and grinds out a table of predictions, the predictor should be chosen at least as carefully as, if not more carefully than, the statistics. In another study, you may wish to cross-validate these predictors. I find this especially important because so frequently the problem arises of comparing regression surfaces from one group to another, and this can be done only if weights for the same orthogonal components are compared.

The only other comment I would make is that I am less pessimistic than Thorndike about the experimental strategy. While telling people to randomize appears to be good advice, I would not condemn out of hand those investigations in which randomization is lacking. This includes, for instance, studies of selected groups, such as pupils in remedial classes, or special cases visited by the social worker. As we look upon the problem as one to be described by a regression surface, then, for the general group, the regression surface is defined ordinarily in terms of the expected achievement. We would expect a particular group receiving special treatment but not necessarily selected at random to depart from their expectancy scores. Now, to be sure, we may not have them located correctly in the predictor space; their true scores will differ from their observed scores on

some predictor variables. Nevertheless, this is a type of investigation we can carry out on a lot of treatments for which we are not ready to invest the efforts of a random sampling study. There is nothing wrong with trying a new cancer cure on hopeless cases and then on people randomly selected. If the selected group responds to this cure it departs from the normal expectancy for people in that stage of disease and can be followed by a random sample study. I think you are a little strong in your condemnation of selective groups because randomization is something that is done only in the later stages of these investigations.

DR. BUNCH: Thank you, Dr. Cronbach, we are now open for comments.

DR. DuBOIS: A point that Cronbach made which I heartily endorse was an implication that in statistics we sometimes reject meaningfulness. In the analysis of training the idea that you can increase meaningfulness by using pairs of defining variables is a profitable approach which has not been sufficiently exploited. Meaningful predictors emerging from replication of relatively pure factor tests have real potentiality in the study of achievement.

DR. JONES: Dr. Cronbach's suggestion that we give structure to our predictors before we use them to predict could be implemented, perhaps, at the item level with the availability of high speed computers. There is a good chance, certainly, that the correlation between two predictors is a function of the item content of the two tests. It might be that you could go down to the base analytical level and come up with different test structures based on different combinations of items.

DR. HARRIS: Most of the formulations we have had today have, tacitly or explicitly, assumed a single criterion. This circumscribes the issue. We have so many variables involved that interest in describing change should not be confined to a single criterion.

DR. CRONBACH: Again, in working with different groups or different treatments, interpretations remain confused because the variables do not remain the same. One can have an educational treatment, at least in principle, that will cause two variables to look very similar when correlated as criteria. Or, there can be two criteria that will be widely separated if a (perhaps) maladroit educational method is used.

DR. TRAVERS: Without a good treatment, there can be no reproduction of the training situation. Maybe the advent of the teaching machine program will provide a reproducible learning situation.

DR. CRONBACH: You agree, then, that no generalization about learning has any meaning without a reproducible treatment?

DR. TRAVERS: I agree with that.

LT. FROEHLICH: I missed the definition of over- and underachievement in both papers. The first paper discusses the investigation in matrix form and the second one has an appraisal definition of underachievement. I wonder, however, in the design of any type of experiment or survey, what finally do we accept as over- and underachievement?

DR. HARRIS: I think Thorndike's advice is, do not do it. Change your strategy.

LT. FROEHLICH: As I interpret Dr. DuBois' comments, you have a matrix of several variables included in which there is an over- or underachievement index.

DR. THORNDIKE: No, you would have an achievement index.

DR. DuBOIS: That is all you would have. I was, of course, going along with Thorndike in saying that you may also be interested in the difference between the final achievement score and predicted achievement. There are two issues involved: one is some sort of a measure to aid individual administrative action, and to provide guidance over the learning process; the other pertains to the general scientific inquiry of the nature of achievement and its correlates.

LT. FROEHLICH: In this morning's second paper, it was noted that the PAQ is based on the grades plus ability measures. If ability is partialed out from the phase grades, the resulting concept is much the same as PAQ. What is the point of determining the correlates of achievement, with ability partialed out, based on the matrix approach?

DR. THORNDIKE: As I understand it, you are using PAQ not in the sense of a predictor, but as sort of a diagnostic tool. However, we were concerned with achievement as the criterion rather than as a predictor or a diagnostic tool. That is, you are using performance in an initial school as one basis for predicting performance in a later school, and, in order to use it more subtly, you are trying to break it into two parts, the part which relates to aptitude, and the part which does not relate to aptitude.

Supposing you have achievement in the preparatory school as one very good predictor of achievement in the later schools. Our interest, then, would be to determine what else can be added that will make it possible to predict any better. If there is nothing you can add, then the concept of under- or overachievement in a class or school no longer means anything.

LT. FROEHLICH: This is very true. However, when used as a predictor, there is little difference between using PAQ in its residual form, as DuBois would have it, or as an achievement index with ability eliminated. My point is that whether under- and overachievement is used as either criterion or predictor, there is little to choose between PAQ and DuBois' matrix.

DR. HARRIS: I would like to eliminate the concept and proceed along other lines. For example, under what conditions can we make any sensible forecasting of various types of achievement?

MR. BERKSHIRE: While we have been referring to that variance which is over-and-above that determined by the predictor as over- or underachievement, if it is systematic it can readily be called under-measure. All this says is that the predictors are not covering the available universe of achievement variance.

DR. CRONBACH: Most people who have dealt with this problem have really assumed that all achievement can be predicted. Thorndike's comments about heterogeneity in areas of measurement are pertinent here. Nobody has done a thorough job of finding out how much we ought to be able to predict. In a sense, that is the hardest part of the research question.

DR. VANDERPLAS: I wonder about the extent to which the definition of over- and underachievement depends on the capitalization of chance phenomena. DuBois said

earlier that if all the predictor variables were available, then the residual score around the regression plane is essentially due to random variation or, at least, error of some sort. It seems to me that the error could be of two kinds: either measurement error within the relatively homogeneous population or error intrinsic to linear correlation. If, for example, achievement is the square function of some predictor, this would be treated as error by the linear regression model, and, we would never be able to find that variance except by experimental techniques which might actually obscure the issue. With the availability of high-speed computers, it might be possible, as Dr. Jones suggests, to go to the item level; but we could also employ higher powers of the variables in order to make the predictions. I would guess that if we had all the variables and all the higher power functions that were available, the remainder would essentially be error of measurement.

DR. BRYAN: In the school or training situation, is administrative action taken on those identified as underachievers, or is such action reserved for those students who are both underachievers and failing?

LT. FROEHLICH: For the high ability underachiever, no action is taken. While there are many students, there are just so many teachers, classrooms and facilities available. As a matter of fact, the night school probably gets only the individual who is failing his courses.

DR. BRYAN: In terms of partialing out ability from the predictors, I was wondering whether achievement qua achievement is, ultimately, the criterion on which administrative action is made. That is, are low achievers dealt with in one way and high achievers in another way regardless of whether they are under- or overachievers?

DR. LORDAHL: In an article concerned with a proposed college system, it was suggested that if the students did not work up to certain levels of predicted capacity, they would be told to take a semester off. Administrative action of this nature could do real damage if it were not completely justified.

MR. BERKSHIRE: We compute the correlations of pre-flight grades with subsequent success in training and determine the appropriate weights for the various factors. We provide the administration with weights that maximize prediction of overall pre-flight grades as well as expectancy tables based on experience with previous students. When a student comes

up for administrative decision as to whether he is to be given another chance, this material is used to tell the administrator what the probabilities of success would be if they retained the man. The information is not provided to help the student.

LT. FROEHLICH: I wonder what Dr. DuBois thinks about a criterion measure of achievement, especially achievement out of which is taken ability, which is probably the major source of variance. But, then we are left with the big problem. If we look for non-intellectual factors to account for the remaining variance in the achievement criterion, what do you suppose we can get with the non-intellectual variables which correlate low with everything else? What is there to partial out which will add to knowledge of what the criterion is?

DR. DuBOIS: The job of the psychologist in learning about complex skills is to understand the correlates of these skills in terms of the relatively invariant characteristics of individuals, teaching methods, personal characteristics of the student, and temporary conditions under which learning takes place.

DR. HARRIS: Would you say that it is important to understand a zero correlation just as much as an r of greater statistical significance? The meaning of a zero correlation could be as important to a good theoretical formulation as positive results.

DR. DuBOIS: I quite agree and, I think the basic word in what you said is "understand." It should be a scientific job. Statistics and engineering are not moral equivalents. However, it seems to me that a multiple correlation does not contribute much to understanding. All it can do is lump predictors together and maximize a relationship with a criterion. Other types of statistical analyses such as partial correlation, part correlation and factor analysis are better tools for trying to understand what is happening in a given situation, and, certainly, much more interesting.

DR. HARRIS: With a simple multiple correlation a linear composite is created which does not necessarily have any operational or rational value, but which still could be used as technical documentation. The particular character of the composite is certainly dependent on a number of chance phenomena as well as on phenomena which the investigator may induce. So, it is a linear composite which has maximized something within that sample, but, beyond this, contributes little to the experimenter's thinking. And this would be a very unsatisfactory place to stop.

DR. CRONBACH: The part correlation is no different from the multiple correlation, for all one does is determined by the multiple correlation between the criterion and another variable with everything else held constant. Hence, whichever partial variance you leave in will be based on a decision just as arbitrary as taking the deviation from the regression line in the first place. Therefore, unless you use a stronger method of analysis, the difficulties encountered with the multiple correlation will also be present in the part correlation. I would like to see the variates reduced to a manageable number, few enough that I could think about them.

DR. DuBOIS: After you have collected your measures and teams of variables which are understandable, perhaps we should start with a test construction program and a criterion construction program in which the units are understandable before we combine them.

DR. CRONBACH: Yes, we would start with about 20 variables, putting them into four blocks. We might think of them as teams and obtain team scores. Then, if we have to make a second-level analysis, which essentially involves differences between teams, we would not have to take that as seriously as the first-level analysis. You could have a part correlation figured at the initial block level, the second block level, and, if you insist, at the individual test score level or, even, the item level. But, some hierarchy of central variables seems necessary to keep the analysis from becoming statistically irrational.

MR. BERKSHIRE: I would like to protest that prediction without understanding is not totally immoral, at least in our business. I can earn a lot of money from the fact that grades in navigation, physical training grades, and peer ratings will consistently combine to give prediction of subsequent failure to the order of .60. While the beta weights vary somewhat, we convert these to practical unit weights anyway. Also, grades in mathematics and physics consistently serve as suppressors. I do not pretend to understand why they are negative correlates, but the correlation is good and we can provide the administrator with probabilities of success or failure in pre-flight training.

DR. HARRIS: I think the point can be illustrated by a judgment of mine in connection with a Ph.D. candidate at Wisconsin. This candidate decided that his doctoral dissertation would be to develop a system for predicting academic success in a particular college. His defense of this

as a desirable topic was that it would be highly useful, since the college needed it. He did not do the dissertation because of my objection that he would simply be utilizing procedures that are well known, making an application to a special case for an essentially utilitarian purpose. This could, in no sense of the word, come into the bailiwick of what the Graduate School calls "some contribution to knowledge." I think all of us are interested in going beyond the concept of this graduate student, but we also recognize the well-documented utility of a pragmatic approach in situations confronting the Air Force and Navy.

DR. MAYO: I would like to see if Dr. Thorndike would be willing to loosen up his fairly rigorous rules for handling experimental groups. At one point in your discussion you mentioned the problem in clinical research was to get equivalent control groups, but in another place you mentioned the role of residual gain. I was wondering whether you actually have to have equivalent control groups in terms of matched scores if you are dealing with gain.

DR. THORNDIKE: The situation in which one treatment has been applied to one group and a different treatment, or no treatment, has been applied to the other group is less critical because the factors associated with the initial status of the groups are most likely to be non-equivalent. An example of the thing I was inveighing against is what happens when a college does research on its remedial reading program. In going through the records of the freshman class, the investigator will find many students whose reading level was as poor as those who took the remedial training, and he may then divide all these students into two groups in order to follow up their subsequent academic success. Now, he often finds that the ones who took the reading program got better grades than those who did not during the following one, two or three semesters.

But, there are several things operating here other than remedial training. What kind of motivational differences exist between the two groups, which lead one to enter the remedial program and the other not to? The reading test is but one fallible measure and the fact of these students being in the reading program or not may well be based on characteristics of the students which pervade their general behavior.

After reviewing some of the underachievement literature, it seems to me that a number of rather dubious experiments which involve the idea of intervention are after-the-fact investigations in which the control groups are either

questionably equivalent to those in the experimental group, or, in one or two other cases, randomly selected before the differential treatment is applied. The results in several instances have been entirely negative.

DR. HARRIS: I would like to make three points. One is that I would strongly urge everyone not to try to match cases. I think there is enough in the literature to document this. The matching business is extremely bad strategy and we ought to wipe it out as rapidly as possible. Secondly, I see nothing wrong with experimental studies incorporating special populations of people, captive groups, etc. Those questions which can be answered satisfactorily at all, will be answered only through replication with other populations, again possibly captive, which nevertheless look like somewhat contrary cases. If it stands up under this kind of replication you have the beginning of a pretty solid generalization. My third point is that you should, if at all possible, make the extra effort to build into the experiment some test of randomization. One can usually, with some arbitrary device, decide who gets into which experimental category.

From the dewey-eyed statistical point of view one would first define an infinite population and then extract a series of random samples from that population to test the hypothesis. This is really quite naive. We are always working within time and space limitations and our populations are always clustered to a considerable extent. If people do not cluster themselves, certainly administrative decisions cluster them. I see nothing wrong with this since we eventually have recourse to replication. If the treatment works on another group that looks as though it would be impervious to the treatment, then the initial hypothesis is strengthened.

DR. DuBOIS: We are often faced with that question at the Memphis Naval Air Station. If we are going to study some of these issues in training, we cannot turn the schools upside down, but must work within the established framework. The general formulation seems to be class-to-class replication. When a new class comes in we can treat it a little differently and, with permission, possibly expose it to the old treatment for comparison. I would think that we get reasonably good replication that way. Then, eventually, we can get generalized knowledge even though we are not set up in a place which affords convenience experimentation.

DR. HARRIS: While this may not be in the elementary books, there are methods for testing clusters. Even though you are forced to work with these pre-arranged

classes, you can decide by some random device which class gets which treatment. And, out of the data, you can then get some pretty decent estimates of error in order to determine whether or not some particular phenomenon is associated with the differences. I do not think the statistics are cramping our style at all. Rather, some of us do not know enough statistics to realize how free and easy we could be on some of these things.

DR. JONES: Some of us are willing to experiment with or without purely random samples, and, obviously, matching techniques have historically been a rational means of control. What do you think about establishing much tighter rules for one of the alternatives to matching, viz., corrections in a co-variance design? In co-variance analysis there is a tendency to seek for variances as something on which to base corrections. This can lead to misleading results.

DR. LORDAHL: From a practical point of view we are looking for some measure that will correlate highly with the ability of the student to benefit from special treatment. If we can increase the probability of the borderline student making it through the course; if we can validate an instrument on which an appropriate decision can be made, then we would have a basis for talking about over- and underachievers. Without empirical evidence, the term, as such, is not meaningful.

DR. DuBOIS: Cronbach has suggested that multiple R's, in predicting achievement, have gone up about .15 in the last 15 years, or a point a year. What are your reasons for this increase? Is there a chance for further increases, or are we now reaching the limit of prediction?

DR. CRONBACH: My best judgment as to the reason is that more time is going into the prediction: more hours of testing and better tests. The increase lies in technically excellent measurements, rather than in the ingenuity in finding new variables. Regarding your second question, my guess is that multiples are not going up much further in predicting averages. I think there is too much variability that comes under the heading of criterion heterogeneity. The only difference between us is that you define the heterogeneity in terms of known sources of variance, whereas I think there is far more variability in the things we do not know about.

DR. DuBOIS: Do you think there is much point in constructing better interest measures, better questionnaires or the situational tests which you said some years ago had not worked out in industry? Does the ingenuity of the test con-

structor hold out very much promise? How about new methods of presentation, such as teaching machines?

DR. CRONBACH: It seems to me this is where the problem ought to be moved. As distinct from predicting a grade average in a heterogeneous training environment, we ought to try to find out more about these interactions. This involves differences in subject matter. As I said earlier, French is getting better results within subject fields and he could probably improve his results if, in addition to content, he standardized the teaching methods within the field. Moreover, we have not begun to consider what variables we can manipulate in the teaching process itself. I would entertain a suspicion that if we used structure as a variable in teaching, the personality variable might well predict response to that, but this is an entirely different line of research.

DR. DuBOIS: Having disposed of over- and underachievement today, our research time should be devoted to ascertaining ways of increasing achievement, evaluating achievement and in developing teaching methods which are related to that achievement.

DR. TRAVERS: I cannot help but feel that some of these predictions are an artifact of the way results are published. Multiple correlations below .60 in the prediction of college grades often do not appear in the literature, while those at the other end of the distribution which are more fortunate in their predictions get published. Hence, we seem to have an upward trend but I am not convinced it is a real thing.

DR. HARRIS: I think that is a good point. I can recall not recommending a particular article for the Journal of Educational Psychology because it was essentially concerned with the usual aptitude test battery predicting performance on a standard achievement test. I can see how I was contributing to possible bias by rejecting a study which yielded correlations of around .45. These are the kinds of relationships we have known about for many years and I do not see that another study of this sort adds anything. However, if studies of this nature are rejected wholesale, this fact would account for the apparent increase in predicting.

DR. TRAVERS: The correlations for the grade criterion have reached .80. If this, in fact, is the probable limit, then motivational variables are useless in achievement prediction.

DR. DuBOIS: In many studies there must be a great deal of spurious overlap between the predictive measures and the achievement measures. We use a 5-alternative, multiple choice format for aptitude tests and tests of reading comprehension, only to use the same format for the final examination. The format may enter significantly into the final grade. We have found that in predicting grades at Washington University, reading comprehension is our best single predictor; and I suspect that reading comprehension has a great deal to do with lots of final grades. I also suspect that our problems may look a little different if college professors ever begin to grade on what students learn in the course, rather than what they know at the end of the course.

DR. CRONBACH: I think a reference to Frederiksen and Melville should be entered in these proceedings. While it does not fit the motivation argument exactly, one of their studies is almost too good to be believed. They obtained two measures of what they thought might be compulsiveness: a difference score between reading speed and comprehension, and high interest scores on the Strong Vocational Interest Blank in areas such as accounting. They found that they could split students on the basis of these indices and found that, for engineers, achievement of compulsive students could be predicted by ability tests, while interest inventories could best predict performance of the non-compulsives. Their results confirmed the hypothesis that non-compulsives work only at what interests them, while compulsive students tend to work hard regardless of their interest in the task.

What these investigators have is a clear indication that the regression surface is tilted, that is, skewed in a space defined by this personality measure, when engineers' grades at Princeton are the criterion. This has been cross-validated with essentially the same results. I think we have been too casual in our discussion in saying one could go into non-linear relations if one wanted to. Here is a strong indication that one ought to want to.

DR. BUNCH: It seems to me that you disposed of the problem of who is an underachiever or an overachiever awfully fast, largely because your predictions were based on something inadequate, rather than because of deficiencies in the concept itself. You may discover that motivation is quite important in the study of underachievement whether the students with whom you are concerned are of high or low ability. You may also discover that if we have some way of increasing motivation, we can increase the level of achievement of students of both high and low ability, that is, throughout the range of abilities.

I would not want to dismiss experimental procedure as something to be avoided in research of this type. Referring to Dr. Thorndike's strategy of prediction over time, what is going to happen in the time interval? One way of finding out would be to manipulate whatever it was that was occurring during the interval, whether remedial treatment or some other variable. I would like Dr. Thorndike to present some of the basic points in the manipulation of appropriate variables.

DR. THORNDIKE: Perhaps we can confine ourselves to the underachievers. Suppose we divide this group and give half of them an intensive course in psychotherapy and see how this affects their grades during the next year. For the other half, perhaps we could have them do calisthenics every day at four o'clock. If you have some reasonable hypothesis for such a procedure and are willing to set up a reasonable experimental design to determine who gets the experimental treatment and who does not, I would have no objection.

However, if we have adequate measures of motivation, then differences in motivation would be as good a predictor for over- and underachievement as other indices, because one could then find the correlation of motivation with achievement. When we determine the correlation of motivation with ability, we must also take steps to extract the non-linear interactions. However, I think most of the things that you can do with the experimental manipulation technique can also be studied within the correlational design of actual prediction over time.

DR. HARRIS: You should be careful to allow for detection, in your analysis, of a possible interaction between the treatment and predictions because you would be dealing with underachievers throughout the ability range. It would be perfectly possible, for example, that no differences be found because the data are obscured by differences within the range. Thus, significant differences may be found in the interactions indicating that the treatments did not help the low ability underachievers, but did help the high ability achievers. The concept of underachievement may mean different things at different points in the scale. That is why I am not content to stop at the point of simply identifying achievers.

DR. BRYAN: Dr. Bunch, were you suggesting that manipulation of Variable X would differentially affect the underachievers as contrasted with the par achievers?

DR. BUNCH: I would suspect it would be a perfectly legitimate hypothesis to say it would affect them all, throughout the range of abilities.

DR. BRYAN: If that is the case, then it would seem to be a matter of straight achievement research, that is, the determination of relationships between predictive variables and achievement. I suggest that the concept of under-achievement is not necessary in order to conduct the kind of investigation you were speaking of when you talked of finding the relationship between a Variable X and underachievement, unless you felt that the manipulation of Variable X would have a differential effect on the underachievers as contrasted with those who were achieving at or above their expected level.

DR. MAYO: In the example of psychotherapy, I think one may be dealing with a non-linear function. Perhaps people who are reasonably well-adjusted will achieve at a satisfactory level, whose those doing poor work in spite of high ability could profit from therapy. It would, presumably, have the effect of raising their level of achievement, but it may have little or no effect on the well-adjusted group. It seems to me it would have a differential effect in many conceivable cases.

MR. BERKSHIRE: If you adopt the experimental approach and set up these treatments, you are accepting the idea that over- and underachievement is actually involved. Yet, we have been in fair agreement that, unless we are sure we have covered the universe of aptitude and ability, we may be dealing only with under-measurement, that is, with some aspect of ability that has not as yet been measured. Because we cannot account for all the systematic variance in the criterion by our predictors, we think of discrepancies in our prediction as over- and underachievement. This would lead us to adopt an entirely different set of hypotheses if we should wish to run an experiment.

MR. SPIES: Given the best measure of aptitude which you can conceive, there is still something over and above that measure which could reasonaly be expected to be predicted by other types of predictors. Since aptitude tests have been used extensively and refined almost to the optimal level, it would appear to be worthwhile to look at these other procedures.

MR. BERKSHIRE: It has already been said that prediction of college achievement has improved in the last ten or fifteen years so that we are now picking up more

systematic variance with our tests. If we had said ten years ago that the variance we could not account for was over- or underachievement, and this probably was done, we would have been wrong because it is now being picked up by other aptitude measures.

DR. THORNDIKE: I would have a good deal of assurance, as long as the measure of achievement is the impression that the individual makes upon his instructor, that there will be something besides aptitude in the criterion. On the other hand, if the measure is his performance on an objective test, I would be less sure. The impression the instructor gains from a student reflects such things as diligence and personal acceptability. The correlation between the aptitude measure and the objective final examination will be higher than the correlation between aptitude and the instructor's impression of the individual at the end of the course. Consequently, we have more variance left over to work with when we are dealing with instructors' impressions.

DR. JONES: You end your paper by saying that the methods to be used in this type of research are largely based on personal choice, yet you obviously recognize at least three approaches, albeit imperfect, which are suitable. I do not think anyone is really in disagreement with either the categories or the designs. Rather, Dr. Bunch is saying there is no particular reason to suggest that the educational situation, itself, should remain static if we want to change it. In order to find out how it can be changed, we need to experiment. We should also recognize that prediction is not just an end in itself. If we find out, with an appropriate measure, that student motivation is poor, perhaps we can improve our predictions. We could also follow Dr. Bunch's suggestion in capitalizing on what we have found and try to produce better achievement by manipulating motivation. These approaches are not really incompatible.

DR. THORNDIKE: Of course, the thing we would really like to do is find intervention in the educational pattern that would affect, not individuals, but the means of achievement so as to push up the whole group.

DR. JONES: Exactly. We would want to especially apply it in those situations where it would salvage the poor achiever. This problem has come up in flight training and has cost the Navy some money. They spend 14 weeks on a man and, when they have to get rid of him, it costs money. Obviously, it would be advantageous to the Navy to boost the fellow up to meet the requirements of the course and, later on, of his job.

DR. THORNDIKE: Unless, of course, he is going to crack up a five million dollar plane.

DR. JONES: Yes. But, while this is a real good reason to get rid of him, it would have been even better not to start him in training. It seems to me there is plenty of room for all these strategies to be used in their appropriate places. The remaining question seems to be where is the appropriate place to use what?

Note: The preceding transcription represents the day's deliberation on the subject of over- and underachievement in academic situations. Appended are Mr. Hackett's bibliography and Dr. Mayo's APA paper.

A Bibliography of Over- and Underachievement in the Classroom

Prepared by:

Edward V. Hackett

1. Ahmann, J. S., Smith, W. L., and Glock, M. D. Predicting academic success in college by means of a study habits and attitude inventory. Educ. Psychol. Measmt., 1958, 18, 853-857.

Over- and underachievers were determined by the deviation from the regression line based on high school average, Co-op Natural Science Test, and the Cornell Mathematics Test. If the student's obtained grade point average exceeded or fell short of the predicted average by at least 4.0 points, he was classified as OA or UA respectively. The amount 4.0 was approximately 2/3 S.D. of the GPA. No significant differences were found between the groups on a study-habits inventory. The inventory added nothing to prediction when added to the test battery and, in fact, did not correlate significantly with grades when used alone.

2. Altus, W. D., A college achiever and non-achiever scale for the MMPI. J. Applied Psychol., 1948, 32, 385-397.

OA and UA were selected by discrepancies of +/- .5 sigmas between standard scores on a measure of verbal aptitude and grade point average in psychology course. An item analysis of the MMPI yielded 60 items supposedly showing trends in immaturity, femininity, and social extroversion for the UA. On the full scale MMPI only Ma scale was significant as determined by mean differences. The 60 item test correlated .39 with achievement and .15 with verbal aptitude. A later analysis of the 60 items with respect to honor point ratio using the upper and lower quartiles successfully discriminated grade averages for the two groups and yielded 26 items. When correlated with GPA and verbal aptitude, the r's were .39 and .31. It should be noted that only 21 of 60 items were significant at the 5% level. In addition, 22 males and 3 females in the OA group versus 9 males and 16 females in the UA group may have loaded the results with feminine interest items which may not hold up under cross-validation. Also, pooled Mf scores are difficult to deal with insofar as a high Mf score for males refers to feminine interests, whereas a high score for women indicates masculine interests.

3. Anderson, Irving H. and Dearborn, W. F. Reading ability as related to college achievement. J. Psychol., 1941, 11, 387-396.

136 Harvard freshman were paired on intelligence but had different scholastic records. Several reading tests were employed to measure possible differences between the 68 pairs. Only the Nelson-Denny Reading Test was found to be effective as a diagnostic tool for college level work. Nevertheless, the authors conclude that reading ability still appears to be related to academic marks.

4. Applezweig, M. H., Moeller, G., and Burdick, H. Multimotive prediction of academic success. Psychol. Rep., 1956, 2, 489-496.

In an analysis of non-intellective factors in college success, academic achievement beyond ability is a function of more than one motivational variable. Variables which do not support academic achievement still may provide a basis for prediction. (In the reviewer's opinion, academic achievement beyond ability requires more than just motivation.)

5. Archambault, R. The concept of need and its relation to certain aspects of education theory. Harvard Educ. Review, 1957, 27, 38-62.

The author examines the ambiguity of the term "need" and analyzes its validity as a hypothetical construct. An alternative explanation based on Allport's functional autonomy concept and Woodworth's "transformation of mechanisms into drives" is favored.

6. Assum, A. L. and Levy, S. J. Comparative study of academic ability and achievement of two groups of college students. J. Educ. Psychol., 1947, 38, 307-310.

This study was designed to determine differences, if any, between adjusted and non-adjusted groups in achievement. Using the ACE (Q, L, Total) and college tests for reading and writing with students in counseling as opposed to non-counseling students, low positive correlations were found between ability and comprehension; no significant differences between groups on ability; adjusted group did better on achievement.

7. Barrett, H. An intensive study of thirty-two gifted children. Personnel Guid. J., 1957, 36, 192-194.

A study of 32 children having Henmon-Nelson I.Q.'s of 130 or more, half of whom were superior students and half of whom were poor students, indicates that the patterns of underachievement and high achievement are apparent by Grade V and continue into secondary school. A number of other conclusions are briefly mentioned.

8. Bendig, A. W. Manifest anxiety and projective and objective measures of need achievement. J. Consult. Psychol., 1957, 21, 354.

The Taylor Manifest Anxiety Scale, McClelland's n-Achievement scale and the achievement scale of the Edwards Personal Preference Schedule were administered to 244 students, 136 males and 108 females. Scores were converted to stanines and r's were computed separately for men and women. There were no significant sex differences. Inter-test correlations indicated that the scales measured different things and also were not related to verbal ability.

9. Bendig, A. W. and Klugh, H. E. A validation of Gough's Hr scale in predicting academic achievement. Educ. Psychol. Measmt., 1956, 16, 516-523.

Both Gough's Hr scale and high-school rank correlated about .32 with GPA. The multiple R for Hr and high-school rank with Quality Point Average was .45. The authors conclude that the Taylor MAS is relatively useless in academic prediction. Although MAS and Hr showed consistent negative correlations, sampling fluctuations prevented MAS from being used as a suppressor variable in predicting QPA.

10. Berdie, R. F. Aptitude, achievement, interest and personality tests. A longitudinal comparison. J. Appl. Psychol., 1955, 39, 103-114.

Thurstone's PMA test, the Strong VIB, Minnesota Personality Inventory and four Cooperative achievement tests were used to determine if special abilities differentiate occupational and educational groups better than interest tests. Students were evaluated at the end of the freshman year in 1939 and followed up in 1949. In terms of curriculum chosen, type of degree and later employment, the vocational interest tests differentiated better. At the college level, differential abilities do not appear to be important as discriminators. Abilities cannot be disregarded, of course, but for counseling purposes, interest tests seem quite valuable. Curriculum choice depends more on motivations and interests than upon special abilities. Of course it might be added that special abilities may reflect themselves as interests and motives.

11. Berger, J. L. and Lutker, A. R. Relationship of emotional adjustment and intellectual capacity to academic achievement in college students. Mental Hygiene, 1956, 40, 65-77.

Entrance and other tests were correlated with academic achievement over a four year period. The authors conclude that students with high intelligence and adequate personality adjustment achieve higher performance. High intelligence with emotional maladjustment invite early attention.

12. Boardman, C. W. and Finch, F. H. The educational and vocational status of University of Minnesota students having low college aptitude rating. J. Educ. Psychol., 1934, 25, 447-458.

119 Students who scored in the lowest 40% of college aptitude rating were observed in terms of their later achievement. 36 college degrees were obtained by students in the group and certain individuals obtained more than one degree. They tended, in general, to be successful in business or politics or the professions. College aptitude rating was considered a poor predictor for individual success.

13. Brooks, M. S. and Weynand, R. S. Interest preferences and academic success. Social Forces, 1954, 32, 281-285.

The purpose of this study was to determine if tests designed for measurement of vocational interest have auxiliary potential for predicting academic success. With 622 students, the Kuder Preference Record correlated with course marks on the average of .49. With ACE partialled out, the average r dropped to .40 or less with most r's around .20.

14. Brown, W. F., Abeles, N. and Iscoe, I. Motivational differences between high and low scholarship college students. J. Educ. Psychol., 1954, 45, 215-223.

97 Dean's List students compared with 46 students on probation within the range of 65-139 on the ACE. 139 was the highest score of the probationers, while 65 represented the lowest score of the honor students. The poor student was characterized as indecisive, procrastinating, unwilling to conform or cooperate. This type of behavior tended to occur outside of class also. With intelligence held constant, interest and motivational factors are seen to contribute to poor scholarship.

15. Brown, W. F. and Holtzman, W. E. A study attitude questionnaire for predicting academic success. J. Educ. Psychol., 1955, 46, 75-84.

An analysis of results obtained from 188 item questionnaire administered to high and low scholarship groups matched on relevant variables, revealed that an inventory heavily loaded with statements referring to attitudes and motivation would be superior to the usual study-habit inventory. A revised inventory correlated .50 (men) and .52 (women) with semester grades.

16. Burgess, Elva. Personality factors of over- and under-achievers. J. Educ. Psychol., 1956, 47, 89-99.

The purpose was to determine whether OA have common personality factors which differentiate them from UA. 492 male freshmen engineering students at Pennsylvania State College were compared to 128 who exceeded or fell below a GPA of 1.33 to .87. The 20 at either extreme called OA or UA. Compared on a battery of personality tests, some differences identified on Rorschach, TAT, PF. OA's less labile, more constricted, emotionally inhibited, intellectually adaptive and controlled. UA's also described.

17. Carter, H. D. Method of learning as a factor in the prediction of school success. J. Psychol., 1948, 26, 249-258.

Author described the development of a study-habits inventory which is a self-report type and points out its possible uses for counseling. The article has much to say about academic achievement being a function of things other than intelligence, especially attitude and study habits.

18. Carter, H. D. Correlation between intelligence tests, study methods tests, and marks in a college course. J. Psychol., 1950, 30, 333-340.

Study habits inventory correlated with achievement measures .33 to .51. With a composite achievement index, .48. In general, a study-habits inventory predicted academic achievement less well for college than high-school. It had lower r than some I.Q. tests; higher than others. Since administration time is short, it may be more efficient as a predictor than intelligence tests. Aspects of study procedure make independent contributions to the prediction of grades.

19. Carter, H. D. The mechanics of study procedure. Calif. J. Educ. Res., 1958, 9, 8-13.

The California Mechanics of Study test was constructed and the 150 self-report items evaluated in terms of discriminating between 200 high achievers and 200 low achievers in grade X. The new test yielded higher correlations with grade point average than did measures of attitudes toward study. $r = .53$ vs. $r = .47$.

20. Clark, J. H. Grade achievement of female college students in relation to non-intellective factors: MMPI items. J. Soc. Psychol., 1953, 37, 275-281.

Two groups of entering freshman girls were administered ACE and MMPI, the ACE scores and later honor point ratio being converted to standard scores. OA and UA determined by the difference between Hr ratio standard scores and ACE standard scores. Mean scores on MMPI categories obtained from both groups and the mean differences computed. Three scales were found significant: D, Ma, K, although this did not hold for both over- and underachiever groups. The author concludes that there is no MMPI profile to differentiate OA and UA.

An achievement scale was then made up of 56 MMPI items which correlated with HPR for group A .28, with ACE -.29. In Group B, scale had $r = .369$ for HPR, and $r = -.303$ with ACE. Conclusion: more reliable results obtainable if the whole range of scores for achievement versus non-achievement dichotomy is used; items for differentiation should have less clinical tenor; and any scale used should be applied only to the sex upon which it has been validated.

21. Clark, R. A., Teevan, R. and Ricuitti, H. N. Hope of success and fear of failure as aspects of need for achievement. J. Abnorm. Soc. Psychol., 1956, 53, 182-186.

In a study of Swarthmore freshmen, students who had high hopes of success or great fear of failure had lower n-achievement scores than middle group.

22. Davids, A. and Eriksen, C. W. The relation of manifest anxiety to association productivity and intellectual attainment. J. Consult. Psychol., 1955, 19, 219-221.

Scores on Taylor MAS were correlated with performance on a 100-word chained association test. Supporting the prediction that anxiety measures drive, significant positive correlations were found between anxiety scores and productivity of associations. High anxious Ss also gave more anxious ideation in associations. Anxiety and productivity seemed to be independent of intelligence as measures by GPA and entrance examinations.

23. DeCarlo, L. M. A comparative study of some characteristics in achievers and non-achievers among children with retarded mental development. Syracuse, N. Y., Syracuse U. Res. Inst., 1957, 197.

Hypotheses formed regarding performance of achiever and non-achiever groups. Thirty-four conclusions.

24. Dilworth, T. A comparison of the Edwards Personal Preference Schedule variables with some aspects of the TAT. J. Consult. Psychol., 1958, 22, 486.

Additional support for the thesis that EPPS scales—especially achievement — and TAT measure different things.

25. Dowd, R. J. Underachieving students of high capacity, J. Higher Educ., 1952, 23, 327-330.

Children who exceed the 90th percentile on ACE but are at or below the 50th percentile on grades classified as underachievers. No differences found for groups on usual indices. Conclusions include: college atmosphere not responsible for underachievement; the same factors operating before college operate in college; paper and pencil personality tests are of little value in differentiating OA and UA; males more prone to underachieve.

26. Drews, E., and Teahan, J. E. Parental attitudes and academic achievement. J. Clin. Psychol., 1957, 13, 328-332.

An attempt was made to determine the attitudes of mothers of high and low achievement students of both gifted and average children in terms of permissiveness, protection and domination. Mothers of high achievers were more authoritarian and restrictive in the treatment of their children than were mothers of low achievers. The parents of gifted high achievers also seemed to have more punitive attitudes with respect to child rearing. Parental attitudes measured by 30 items from Shoben's 85-item scale. S's were 63 junior high school students divided equally as achievers and non-achievers. No difference between groups on socio-economic level of families.

27. Duff, O. L. and Siegel, L. Biographical factors associated with academic over- and underachievement. J. Educ. Psychol., 1960, 51, 43-46.

This study suggests that many studies in this area really involve a comparison between high and low ability students since high ability students tend to be underachievers and low ability Ss overachievers. Meaningful inferences must avoid criterion construction at either end of the continuum and refer all results to the ability level of the subjects.

28. Easton, Judith. Some personality traits of underachieving and achieving high-school students of superior ability. Bull. Maritime Psychol. Assn., 1959, 8, 34-39.

Four hypotheses were presented which would differentiate over/underachievers:

H1: underachievers of superior ability show less satisfactory parental relationships.

H2: UA have more insecurity and felt inferiority.

H3: UA have more egocentricity.

H4: UA have less achievement drive.

Hypotheses 1, 3, 4 were sustained using two groups of 20 each and California Test of Personality, Thurstone Interest Schedule, a questionnaire and TAT. The first two tests were most discriminating.

29. Ferguson, G. A. On learning and human ability. Canad. J. Psychol., 1954, 8, 95-112.

In an attempt to draw together crudely within the same scheme both the study of learning and the study of human ability, several views are advanced: an individual may possess the ability to perform a task adequately but may lack the ability to perform the task under particular learning conditions; those abilities which are important to survival in a given culture will increase with age; the correlations amongabilities are explained in terms of positive transfer, and their differentiation by the development of abilities specific to particular learning situations.

30. Fliegler, L. A. Understanding the underachieving child. Psychol. Rep., 1957, 3, 533-536.

The underachiever may be a maladjusted child whose primary difficulty stems from inadequate home or school relationships. Distorted interpersonal relationships lead to negativistic teacher identification and an inability to achieve. Specifically, lowered aspiration and frustration have to be considered. A discussion of problems and requirements for counseling the underachiever is included.

31. Frankel, E. A comparative study of achieving and underachieving high-school boys with high intellectual ability. J. Educ. Res., 1960, 53, 172-180.

Equivalent groups matched on age, I.Q., entrance score, race; separated on GPA. Achievers superior in math and verbal aptitude--especially math; achievers interested in science, underachievers interested in artistic-mechanical areas; achievers tend to think about the future, underachievers about present problems. Achievers missed half as many days from illness as underachievers, although both same in health. Achievers had more professional parents--more education.

32. French, Eliz. The interaction of achievement motivation and ability in problem solving success. J. Abnorm. Soc. Psychol., 1958, 57, 306-309.

To test the relationships among n-Ach, ability level and problem-solving behavior, 96 airmen in basic training were divided into three groups on the basis of the Armed Forces Qualification Test. Achievement motivation was measured with the Test of Insight by scoring the achievement oriented explanations for the hypotheses formed in the solution of the problem situations. The hypotheses that intelligence is more related to success when n-Ach is high and that high n-Ach persons solve more problems than low n-Ach types were sustained.

33. Garrett, H. E. A review and interpretation of investigations of factors related to scholastic success in colleges of arts and sciences and teachers colleges. J. Exp. Educ., 1949, 18, 91-138.

A review of 194 studies related to scholastic success at the college level. Measures having the greatest prediction value are, in descending order of correlation: high-school scholarship, general achievement tests, intelligence tests,

general college aptitude tests, special aptitude tests. Multiple R's of two factors usually result in a somewhat higher correlation with a criterion than do the factors taken singly. Addition of a third factor adds very little. No personality test yet devised can predict, to any appreciable extent, college success.

34. Gebhart, G. G. and Hoyt, D. P. Personality needs of under- and overachieving freshmen. J. Appl. Psychol., 1958, 42, 125-128.

In order to study personality correlates of OA and UA while controlling variables which earlier studies had neglected, Ss were matched for sex, level of academic progress, and different ability levels. Also, the article criticizes the use of different personality tests to measure correlates, and the vague definitions of OA and UA. This study attempts to see if personality correlates maintain themselves at different ability levels and within different vocational objectives. Ss scoring above or below predicted GPA were classified accordingly. Cut-off score was above and below .70 to 1.30 GPA using a 3 point grading system. Overachievers were high on n-Ach, n-Ord, m-intraception, m-consistency. Underachievers were high on n-Change, n-Affiliation and n-Nurturance. High ability students were higher on need-achievement, exhibition, autonomy, dominance, consistency; they were lower on needs-deference, order, abasement, nurturance. Engineering students were higher than Arts and Science students on endurance; lower on dominance.

35. Gough, H. G. The relationship of socio-economic status to personality inventory and achievement. J. Educ. Psychol., 1946, 37, 527-540.

The correlations between socio-economic status and achievement averaged .30, except for achievement in arithmetic which correlates .07. Personality test and achievement yield r's from -.22 to -.28. With intelligence partialled out, r's are somewhat less. Among other conclusions, Gough suggests that socio-economic status has slight positive relationship with academic achievement. Personality tests yield slight negative correlations.

36. Gough, H. G. Factors relating to the academic achievement of high-school students. J. Educ. Psychol., 1949, 40, 65-78.

One reason why studies fail to correlate is the use of scale scores from instruments which were devised for use in other prediction problems - often clinical with no intended relationship with the variables relevant to OA and UA. Thus, there is no reason to believe that a scale designed to test neuroticism would discriminate OA and UA as a group. In reviewing earlier studies, Gough suggests that while introversion, dominance, self-sufficiency, motivation, liberal social attitudes and lack of maladjustment are characteristics of achievers, curricular satisfaction, maturity of goals, efficiency of planning and working (study habits) and adequate personal and social orientations are involved. On MMPI, high scores on Ma, Pd, and Pt may mean underachiever type. Similarly, items which reflect a lack of emotional tension, immaturity, social extroversion, a disinclination to admit personal problems and tendency to see others in a favorable light tend to "predict" underachievement. In this study, one of the MMPI scales differentiated OA from UA for high-school students. May mean a need for a wider range of items than that of MMPI. (Also, clinical tests usually designed with concurrent and construct validity which may not predict well.)

37. Gowan, J. C. The underachieving gifted child. Exceptional Children., 1955, 21, 247-249.

Directed toward teachers and counselors, the article classifies those characteristics of gifted children who are presented for educational counseling for poor to failing work. These include familial (parents) problems and their ramifications, ease of school work, etc. Also, self-sufficient and unsociable child identifies less with parents and finds it difficult to find surrogate in teachers who usually are overachievers themselves.

38. Gowan, J. C. Dynamics of the underachievement of gifted children. Exceptional Children, 1957, 24, 98-101.

A review of the literature shows the underachieving gifted child has poor ego controls, little definition of school or occupational choice or academic goals, autocratic or laissez-faire parents. They tend to a kind of intellectual delinquent who withdraws from goals, activities, and active social participation in general.

39. Gowan, J. C. Intelligence, interests, and reading ability in relation to scholastic achievement. Psychol. Newsltr., N.Y.U., 1957, 8, 85-87.

The study indicates that scholastic achievement correlates highest with reading ability, less with intelligence, and least with interests.

40. Grooms, R. R. and Endler, N. S. The effect of anxiety on academic achievement. J. Educ. Psychol., 1960, 51, 299-304.

In studying anxiety and achievement, the authors conclude that separate regression equations be established for high-anxious subjects. The over-all correlation between anxiety scores and grades was .30, but when the group was trichotomized, r's were .63, .13, and .19 for HA, MA, and LA respectively.

41. Harris, D. The relation to college grades of some e factors other than intelligence. Arch. of Psychol., 1931, No. 131, p. 55. (147 refs.)

Intelligence must admit other bedfellows in predicting grades in college. In an analysis of 456 CCNY freshmen matched on cultural bases, scores from a variety of intellectual, personality, health, and interests examinations indicated that students receiving lower grades than expected from their Alpha scores were: non-conforming in religious and other areas, extroverted, preferred English, rejected mathematics, and came to college more for prestige than knowledge. Non-discriminating factors included: age, physical details, number and kind of books read, and lack of vocational choice. Provides a review of over/underachievement literature prior to 1930.

42. Harris, D. Factors affecting college grades: a review of the literature, 1930-1937. Psychol. Bull., 1940, 37, 125-166. (328 references)

The author presents a detailed classification and summary of the results of studies (primarily correlational)of relationships between college grades and intelligence, high-school grades, physical data, personality, interests, attitudes, non-grade high-school factors, study habits, teaching methods and conditions, incentives and direct motivation, amount of course work taken, curricula and occupational choice, and extra-curricular factors. Methodological faults in such studies are discussed. Intelligence still best single factor in predicting grades (r's = .33 to .64). Some find high-school best (r's = .60 to .78) especially in predicting failure. In a concluding statement, the author felt that many studies indicate that the essential factors in student achievement are: 1) ability (I.Q., scholastic aptitude), 2) effort (drive, motive), 3) circumstances (social, economic, academic).

43. Hewer, Vivian H. A comparison of successful and unsuccessful students in the medical school at the University of Minnesota. J. Appl. Psychol., 1956, 40, 164-168.

Unsuccessful medical students had a significantly higher L score on MMPI than did the successful candidates. The Strong VIB was of no value. The rater could make no use of either MMPI or Strong to identify successful and unsuccessful students.

44. Himelstein, P., Eschenbach, A. E., and Carp, A. Inter-relationships among three measures of need achievement. J. Consult. Psychol., 1958, 22, 451-2.

A comparison of 298 Air Force Academy freshmen of the n-Ach scale on the EPPS, McClelland's n-Ach test, and French's Test of Insight revealed no significant intercorrelations.

45. Hopkins, J., Mallison, N, and Sarnoff, I. Some non-intellectual correlates of success and failure among university students. Brit. J. Educ. Psychol., 1958, 28, 25-36.

Men and women college graduates were compared with students who either had failed or withdrawn for academic reasons, on responses to a 63 item questionnaire covering social, educational, and economic background; health; pre-university orientation; attitudes toward university life and study. Differentiating variables included: type of school previously attended, scholarship aid, parents' education, and personal-social relationships and motivation.

46. Hoyt, D. P. and Norman, W. T. Adjustment and academic predictability. J. Counsel. Psychol., 1954, 1, 96-99.

Tests the hypothesis that since maladjusted persons may overcompensate and do better in performance or may dwell so much on their problems that they do little, academic ability tests would predict less well for the maladjusted group than for a well-adjusted group. Maladjustment defined as T = 70 or more on two or more MMPI scales (excluding Mf); "one-peak" group had scale value of T $\geq$ 70 on one scale (excluding Mf); normal had no values over 60. Hypothesis confirmed with r's for maladjusted group grades and Ohio Psychological Exam equal to .31, while for the normal group r = .62. An analysis of over/underachievers indicated that: 1) the need for

separate regression equations for normal and maladjusted entering students, and 2) counselors should note the size of the correlation between ability and achievement in evaluating counselee insofar as normals tend to have higher r's.

47. Jackson, R. A. Prediction of academic success of college freshmen. J. Educ. Psychol., 1955, 46, 296-301.

Among other findings, the author concludes that: 1) reading tests are the best predictors of academic success, 2) women obtain higher grades than men, 3) women perform more nearly to their abilities than men.

48. Josephina, Sr. CSJ. Reading accomplishment of gifted and average pupils. Educ. Psychol. Measmt., 1958, 18, 867-871.

Comparing gifted pupils from the 5th and 6th grades (mean I.Q. of 137) with average students from same grade (mean I.Q. of 101) in terms of discrepancy scores for expected and obtained averages in reading vocabulary and comprehension, the average students had lower discrepancy scores. Bright students were considered "retarded" relatively speaking inasmuch as they seldom reached their predicted scores.

49. Kahn, H. and Singer, E. An investigation of some of the factors related to success or failure of school of commerce students. J. Educ. Psychol., 1949, 40, 107-117.

Two groups of upperclassmen in commerce school selected on the basis of being on Dean's list or probation were compared on performance on MMPI and Rorschach. Only measures of personality adjustment could account for the fact that students of high ability fail and some of low ability gain success.

50. Karlan, S. C. Failure in secondary school as a mental hygiene problem. Mental Hygiene., 1934, 18, 611-620.

Thirty-one high I.Q. students with failing grades were classified according to timidity and inferiority feelings or emotional immaturity in accounting for the failure.

51. Karolchuck, P. A. and Worell, L. Achievement motivation and learning. J. Abnorm. Soc. Psychol., 1956, 53, 255-257.

Replicating Lowell's study (58), the authors hypothesized that high n-Ach is related to greater learning in both directed and incidental learning situations. Need-Achievement measured by McClelland's test for 108 students. While the major hypothesis was not supported, the high n-Ach students did show more efficient learning in incidental material. The achievement need index was considered to be unclear.

52. Keys, Noel, and Whiteside, G. H. The relation of nervous-emotional stability to educational achievement. J. Educ. Psychol., 1930, 21, 429-441.

High correlations were found between the variables: industry, I.Q., perseverance, dependability, and ambition as measured by teachers' ratings and grades. Pupils with emotional difficulties tend to do less well in school having achievement retardation averaging one year. The authors suggest that the accomplishment quotient not be used to compare relative educational achievement in groups which differ widely in I.Q. and C.A.

53. Kimball, Barbara. The sentence-completion technique in a study of scholastic underachievement. J. Consult. Psychol., 1952, 16, 353-358.

Twenty subjects scoring high on I.Q. tests but failing in high-school were given sentence-completion tests to determine if underachievers have more negative responses to their fathers, and if such aggressive feelings are a source of guilt and anxiety more often in the underachiever because of an inability to give vent to these feelings. The results seemed to confirm both hypotheses. Advantages and disadvantages to the sentence-completion method in this context are discussed.

54. Kirk, Barbara. Tests versus academic performance in malfunctioning students. J. Consult. Psychol., 1952, 16, 213-216.

Attempts to describe the basic symptomatology when the discrepancy between achievement and capacity is severe and chronic, and to depict the related problems of measurement of capacity in students who are not performing well academically, and whose test performance would therefore be clinically suspect. She criticizes the notion that individual tests measure capacity while paper and pencil types test actual performance which is more likely to be like academic behavior. When capacity is not in doubt, S usually performs well on all tests--aptitude or capacity. Several case histories of underachievers are presented and discussed.

55. Klausmeir, H. J. Physical, behavorial, and other characteristics of high and low achieving children in favored environments. J. Educ. Res., 1958, 51, 573-581.

High achieving children in 3rd and 5th grades were not significantly different in height, weight, strength of grip, permanent teeth or carpal age. High achievers were superior in mental age, occupational level of parent, and classroom conduct. More girls than boys in high achiever group.

56. Krug, R. E. Over- and underachievement and the Edwards Personal Preference Schedule. J. Appl. Psychol., 1959, 43, 133-136.

In a replication of the Gebhart-Hoyt study (34), Krug attempts to answer the questions of whether the PPS can differentiate between OA's and UA's, and whether an aptitude battery compares well with three achievement tests and high-school rank as a basis for determining over- and underachievement. Two samples, each with 120 S's, were selected and termed "aptitude" or "performance" based. Twenty students were retained at each of three levels of expected performance for both over- and underachievement groups. Differences between groups were found according to the EPPS. "For purposes of selection, the EPPS and certain evidences of past performance are functionally equivalent." Krug agrees with Gebhart and Hoyt that there may be several patterns of OA and UA.

57. Levy, N. M. and Cuddy, J. M. Concept learning in the educationally retarded child or normal intelligence. J. Consult. Psychol., 1956, 20, 445-448.

23 pairs of 4th graders matched for age, sex, socio-economic status were places in a concept-learning task. CMM I.Q.'s ranged from 98 to 103. One group was working up to grade placement, others were behind one-half to two and one-half years. The normal achiever made fewer errors than underachievers.

58. Lowell, E. L. The effect of need for achievement on learning and speed of performance. J. Psychol., 1952, 33, 31-40.

McClelland's TAT n-Ach test could predict degree of learning. The group with a high n-Ach increased its output more from beginning to end of a scrambled words task than the low n-Ach group.

59. Lum, Mabel K. M. A comparison of under- and over-achiever female college students. J. Educ. Psychol., 1960, 51, 109-114.

Using SSHA and sentence-completion form, the author concludes that the UA procrastinates, has less self-confidence, requires external pressure to perform (has not "internalized" achievement drive), is more critical of school and of the prevailing philosophy of education. No significant differences were found in actual study habits.

60. Malloy, J. An investigation of scholastic over- and underachievement among female college freshmen. J. Couns. Psychol., 1954, 1, 260-263.

Over- and underachievers defined as top and bottom 27% of students deviating from regression line based on ACE-L score and the Nebraska English Achievement Test score. R with grades for these tests was .587. Characteristics of over- and underachievers based on a Life Experience Inventory are discussed. The Inventory yielded 64 out of 201 items which differentiated groups.

61. Malloy, J. The prediction of college achievement with the Life Experience Inventory. Educ. Psychol. Measmt., 1955, 15, 170-180.

Further validating information on the LEI.

62. Martire, J. G. Relationships between the self-concept and differences in the strength and generality of achievement motivation. J. Pers., 1956, 24, 364-375.

Students with high n-Ach scores had significantly greater discrepancies between self-ideal and self-ratings.

63. McCurdy, H. G. Basal metabolism and academic performance in a sample of college women. J. Educ. Psychol., 1947, 38, 363-372.

With N = 30, correlations were found for BMR and Otis scores, grades, age. Correlation for BMR and HPR was .24, significant at 5% level. The interaction of Otis scores and BMR accounted for over 50% of variance in grades.

64. McQuarry, J. P. Some relationships between non-intellectual characteristics and academic achievement. J. Educ. Psychol., 1953, 44, 215-228.

Factor analysis of non-intellective variables in college achievement yielded 7 factors including: socio-economic-cultural factors, high-school rank -- paper/pencil tests performance, hours studied, etc.

65. McQuarry, J. P. Differences between over- and under achievers. Educ. Adm. Superv., 1954, 40, 117-120.

OA/UA defined as 1/2 sigma above or below the standard score on ACE compared with the standard score on GPA. Students compared on biographical data, interests, high-school activities and parental background. Characteristics of both groups discussed in detail.

66. McQuarry, J. P. and Truax, W. E., Jr. An under-achievement scale. J. Educ. Res., 1955, 48, 393-399.

Defining over- and underachievers in the same manner as given supra, 50 males divided, with 27 in OA group, 23 in UA group. Twenty-four MMPI items differentiated the groups (CR = 2.33 or above). Under cross-validation, the new group was correctly identified on 50% of the cases. Controlling for intelligence by taking only underachievers scoring over the 40 percentile on ACE and overachievers scoring less than 60 percentile on ACE, prediction rose to 77.2% for UA's and to 90.9% for OA's. It was suggested that entrance examinations be cued to admitting lower ability students with overachiever characteristics.

67. Molton, R. S. Differentiation of successful and unsuccessful premedical students. J. Appl. Psychol., 1955, 39, 397-400.

Non-intellective factors such as interests, socio-economic status were of no consequence in predicting performance. Mult. R for high-school performance, ACE, and Coop English Test with grades was .65.

68. Miller, K. S. and Worchel, P. The effects of need achievement and self-ideal discrepancy on performance under stress. J. Pers., 1956, 25, 176-190.

Under conditions of threat to self-esteem (repeated failure) a curvilinear relationship between one's evaluation of inadequacy in coping with frustration and efficiency in maintaining accuracy of performance obtains. S's who evaluate themselves as slightly inadequate relative to the level of expectancy are more efficient in terms of accuracy of performance than those with a self-evaluation of either high or low adequacy.

69. Morgan, H. H. A psychometric comparison of achieving and non-achieving college students of high ability. J. Consult. Psychol., 1952, 16, 292-298.

132 male sophomores were divided by HPR into 3 groups (high, middle, low). While all S's had ACE scores at or above 136, there were significant differences among them in high-school standing. All were administered the Strong VIB, TAT, MMPI, and a questionnaire. Among other conclusions, achievers tend to have: seriousness of interests and maturity; awareness of and concern for others; a sense of responsibility; dominance, persuasiveness and self-confidence; motivation to achieve. It was also noted that n-Ach as defined by TAT correlates with actual achievement only at the upper level with any significance.

70. Munger, P. F. and Golckerman, R. W. Collegiate persistence of upper and lower 1/3 high-school graduates. J. Couns. Psychol., 1955, 2, 142-145.

The persistence in college students who ranked in either the upper or lower 1/3 of their high-school graduating class was investigated by correlating college grades and scholastic aptitude scores. Subjects were placed in "persistence" groups according to the number of semesters for which they had enrolled and the point at which they withdrew or graduated. A significant correlation was found between 1st semester grades and persistence, and persistence was significantly different for students in either third.

71. Myers, R.C. and Schultz, D. G. Predicting academic achievement with a new attitude-interest questionnaire. Educ. Psychol. Measmt., 1950, 10, 654-63.

A small positive correlation was obtained between an interest questionnaire and an achievement index. The index was a regression equation resulting from a comparison of the mathematical and verbal sections of the SCAT with 1st year GPA. The mean of the index was 13 with an S.D. of 4. Those falling "farthest" from the regression were underachievers, those "farthest" above were overachievers. The top 37 students were

used, finally, vis-a-vis the bottom 37. Overachievers were also overachievers in high-school and while their high-school grades were higher than were those of the underachievers, their aptitude scores were lower.

72. Myers, R. C. and Schultz, D. G. Predicting academic achievement with a new attitude-interest questionnaire: II. Educ. Psychol. Measmt., 1953, 13, 54-64.

Supplementary data on the questionnaire previously described. Multiple R remains at .64.

73. Neel, M. O. and Mathews, C. C. Needs of superior students. J. Higher Educ., 1935, 6, 29-34.

Non-achievers of high I.Q. and low grades had religious and life conflicts more than achievers of equal intelligence with higher grades. The former also had more erratic study habits.

74. Owens, W. A. and Johnson, W. E. Some measured personality traits of college underachievers. J. Educ. Psychol., 1949, 40, 41-46.

Using deviations from ACE predicted grades, three groups (over-, under-, and normal-achievers) established and administered 300 MMPI items. 38 items differentiated both over- and underachievers from normal-achievers. Underachievers were more socially oriented, probably too active, had slight tendency toward depression, worry, Psychic tension was product of poor achievement rather than its cause.

75. Parrish, J. and Rethlingshafer, D. A study of the need to achieve in college achievers and non-achievers. J. Gen. Psychol., 1954, 50, 209-226.

48 college students who differed in achievement but who were matched in other respects were selected to test the hypothesis that high and low achievers in equated groups should reveal differences in n-Ach scores of the McClelland type. The hypothesis was not confirmed.

76. Robinowitz, Ralph. Attributes of pupils achieving beyond their level of expectancy. J. Pers., 1956, 24, 308-317.

This study undertook to examine ways in which a given group of overachievers differs from three control groups. Students were selected on the basis of Tsao's "effort quotient" as an index of relative achievement. Nine 11th graders whose EQ exceeded the mean of the class by at least 1.5 sigmas were compared with: a group with average ability and achievement, one with high ability and high achievement, and one with high ability and average achievement. Q-sort, TAT, and the Social Distance Test were used to assess attitudes toward: 1) acceptance by peers, 2) acceptance by family, 3) self-concept, 4) school achievement. None of the tests differetiated the groups, although a trend of ambivalence toward family and peer acceptance was interpreted as leading to compensatory behavior by the experimental group.

77. Schutter, G. and Maher, H. Predicting grade-point-average with a forced-choice study activity questionnaire. J. Appl. Psychol., 1956, 40, 253-257.

In an attempt to introduce the lesser transparency of the forced-choice technique into the study test area, preference and discrimination indices were computed from responses of 99 over- and underachievers to 300 attitude, skill and unclassified items. Cross-validation on 100 students yielded an r = .36, with a corrected odd-even reliability of .83. The forced-choice method in general and the discriminating items for over- and underachievers are discussed.

78. Segel, D. and Gerherich, J. R. Differential college achievement predicted by ACE. J. Appl. Psychol., 1933, 17, 637-645.

For best over-all prediction, general achievement tests are best (median r is .545), then general mental tests (median r is .44), followed by specific tests of various sorts (median r is .37). Contains over 100 references.

79. Shaw, M. C. and Brown, D. J. Scholastic underachievement of bright college students. Personnel Guid. J., 1957, 36, 195-199.

28 underachieving bright students compared with 30 overachievers. While achievement and ability tests yielded no differences between groups, a significant difference did obtain for grade averages. The underachieving students appeared to have and express more hostility toward others, especially authority figures. Underachievement was considered to be related to the individual's basic personality matrix.

80. Shaw, M. C. and Grubb, J. Hostility and able high-school underachievers. J. Couns. Psychol., 1958, 5, 263-266.

To verify the hypothesis that underachievers tend to be more hostile than their achieving counterparts, three hostility scales were administered to a group of high-school male students. Both groups were matched on intelligence and all were above average intellectually. The bright underachievers scored significantly higher on the scales than did the bright achievers.

81. Shaw, M. C. and McCuen, J. T. The onset of academic underachievement in bright children. J. Educ. Psychol., 1960, 51, 103-108.

Achievement and underachievement in children with I.Q.'s over 110 were compared at each grade level from 1 through 11. For males, the underachievers got lower grades than achievers at grade 1 and significantly lower at grade 3. Female underachievers tended to do better than achievers for grades 1 through 5, then dropped off sharply after grade 6.

82. Sperry, B., Staver, N., Reiner, B. S. and Ulrich, D. Renunciation and denial in learning difficulties. Amer. J. Orthopsychiat., 1958, 28, 99-111.

Seven unaggressive, compliant boys with school difficulties are described as renunciating success. They have responded to a family pattern in which they can derive some dependent satisfactions from school failure. Although families on the surface present a picture of success and stability, many compromises and sacrifices have led to a masochistic pattern of relationships. Despite their efforts to dissuade, their own patterns and unfortunate events have persuaded one of their children that he can succeed only by failing. One child in family usually is aggressive and family releases hostile feeling through him vicariously.

83. Stagner, R. The relation of personality to academic aptitude and achievement. J. Educ. Res., 1933, 26, 648-660.

The linear relationships among intelligence, achievement and personality measures are low and probably are so because of the nature of the variables. Extreme personality trends may counterbalance advantages in aptitude making for equal achievement in opposed groups. High emotionality and

self-sufficiency lead to lower achievement than predictable from I. Q. scores. Personality factors have a marked influence on the correlation between aptitude and achievement. The poor prediction of grades by I.Q. scores obtain because of defects in methods of grading, and variable energy levels of the students. Arguments for the use of personality measures for improved counseling are presented.

84, Stephenson, W. The prior analysis of questionnaires. In The Study of Behavior., The University of Chicago Press: Chicago, 1953, Chap. IX 190-218.

Stephenson presents the Q-technique in an analysis of over- and underachievers and contrasts this approach to the usual use of questionnaires which differentiate subject responses according to mean differences. Data from 20 subjects analyzed along Q-sort lines are compared to that obtained from 150 pairs of Ss using more customary techniques.

85. Stogdill, R. M. Personal factors associated with leadership: a survey of the literature. J. Psychol., 1948, 25, 35-71.

Stogdill summarizes the results of 124 studies on leadership factors into the following: above average capacity, achievement, responsibility, participation in events, desire for status, and the effects of the situation on the emergence of a leader.

86. Stone, D. and Ganung, G. A. A study of scholastic achievement related to personality as measured by the MMPI. J. Educ. Res., 1956, 50, 155-156.

Students scoring T = 70 on one or more scales of the MMPI had lower grades than students within the "normal" range.

87. Taylor, J. and Spence, K. M. The relationship of anxiety level to performance in serial learning. J. Exp. Psychol., 1952, 44, 61-64.

Anxious (high drive) and non-anxious (low drive) subjects as determined by scores on the MAS were compared on errors and trials in a serial learning situation. The high D subjects made significantly more errors and took more trials than the low D Ss. Also, since the stimulus was essentially a trial and error task, those conditions which produced the greatest number of competing alternatives differentiated the high D and low D the most.

88. Teahan, J. E. Future time perspective, optimism and academic achievement. J. Abnorm. Soc. Psychol., 1958, 57, 379-380.

TAT stories scored for 60 seventh and eighth grade boys, 30 of whom were in the top quarter of their class (achievers), 30 of whom were in the bottom 25% (non-achievers). While no attempt was made to control for intelligence specifically, the correlation between the time perspective measures and I.Q. was zero. The mean I.Q. for the achieving group was 108, for the low group, 95.5. High achievers gave stories yielding more themes of greater optimism, long-range foresightedness and projection into the future. Low achievers orient themselves largely in terms of the present.

89. Tiebout, H. M. The misnamed "lazy" student. Educ. Record., 1943, 24, 113-129.

In a three year clinical study of girls at Sarah Lawrence College, the "lazy" student was characteristic of students whose scholastic records were poorer than aptitude tests would predict. This consisted mainly of personality characteristics, viz., a need to rely upon strong and immediate motivation to start studying; a tendency to have interests of a transitory nature; a tendency to be governed by strong hedonistic principles; and a deep-seated problem in learning. In spite of a superficiality and disorganized quality pervading written work, a tendency to gloss over failures and to rationalize poor achievement, the "lazy" student shows continued optimism about changing for the better.

90. Turney, A. H. Intelligence, motivation, and achievement. J. Educ. Psychol., 1931, 22, 426-434.

The article suggests that if a "g" factor is present in intelligence, it may not reflect itself in all conditions or circumstances, i.e., not all classroom activities would correlate highly with "g". Many would be a function of industry, perseverance, dependability and ambition. Motivation should be brought into consideration when relating ability and performance.

91. Uhlinger, C. A. and Stephens, M. W. Relation of achievement motivation to academic achievement in students of superior ability. J. Educ. Psychol., 1960, 51, 259-266.

Ss were 72 special merit scholarship freshmen, reasonably homogeneous as to pertinent variables. Slight support was found for the H: that high achievers would have high n-Ach. High achievers tended to have greater needs for social love and affection, relative to recognition, than low achievers. Also, they had a greater expectancy of success and a higher level of aspiration. The low inter-test r's suggested important inadequacies in the concept and measures of need-achievement.

92. Wedemeyer, C. A. Gifted achievers and non-achievers. J. Higher Educ., 1953, 24, 25-30.

Students with I.Q.'s over 130 were divided into achiever and non-achiever groups on the basis of distinguished merit in grades or leadership (awards, scholarships, prizes, etc,) Under-achievers got average or above marks but no distinctions; some were on probation. Achievers were superior in reading (vocabulary and comprehension). The author provides some general comments on the waste of it all and gives suggestions as to what the schools should do.

93. Weigand, G. Adaptiveness and the role of parents in academic success. Personnel Guid. J., 1957, 35, 518-522.

The successful student has been taught to behave as an adaptive individual in all situations. This behavior is supported by the parents' attitude.

94. Weiss, P., Wertheimer, M., and Groesbeck, B. Achievement motivation, academic aptitude and college grades. Educ. Psychol. Measmt., 1959, 19, 663-666.

Comparing the TAT and EPPS, the authors found an r of .26 significant at 5% level. Multiple R for TAT, PPS and academic aptitude test with GPA is .68 which is significant beyond .01 level. Authors conclude that the PPS and TAT may be measuring the same thing but in different ways. Also, the use of females in previous studies may have reduced the correlation between the two. This study used 60 male sophomores.

95. Williams, J. E. Modes of failure, interference tendencies and achievement imagery. J. Abnorm. Soc. Psychol., 1955, 52, 573-580.

This study reports the effects of failure to reach goals set by the subject as opposed to those set by the examiner. While no differences in the effects of the two failure

procedures obtained, the high achievement imagers showed greater improvement following failure to reach self-set goals. The low achiever showed improvement after failing examiner-set goals. High achievement imager worked significantly faster on tasks.

96. Wolf, S. J. Historic background of the study of personality as it relates to success or failure in academic achievement. J. Gen. Psychol., 1938, 19, 417-436.

The author presents an annotated review of the literature for studies of personality factors in over/underachievement prior to 1938. Bibliography contains 81 entries.

97. Wolking, W. D. Predicting academic achievement with the Differential Aptitude and the Primary Mental Abilities Tests. J. Appl. Psychol., 1955, 39, 115-118.

S's were 139 girls and 128 boys in eleventh grade. A counter-balanced design was used to offset practice effects. The DAT shows higher validities than PMA although there is a moderate to substantial relationship between the subtests of the two batteries. The tests do not generally predict best in the subject usually assumed to be measured by the test. All tests show their greatest effectiveness in the prediction of science, geometry and algebra. The author notes that school grades are only one criterion by which to judge the usefulness of these tests, since they are designed for prediction of multiple criteria.

98. Wrenn, C. G. and Humber, W. J. Study habits associated with high and low scholarship. J. Educ. Psychol., 1941, 32, 611-616.

Students paired on ACE, academic major, academic load, academic experience, lack of training in study habits and sex. One member of each pair was in the lower quarter of the class, the other was in the top 20%. The Wrenn Study Habits Inventory was sent to all pairs by mail and 89% responded. Men appear to be more variable in study habits as a source of academic variability than do women. The criteria for the high-low groups were the Dean's List and the probation list.

99. Wright, S. Some personality characteristics of academic underachievers. Abstract, Amer. Psychol., 1954, 9, 496.

A comparison of two groups paired on relevant variables by means of a modified covariance technique which adjusted for poor matching in three variables. Pairs exposed to a battery of tests consisting of MMPI, several projectives and GSR. Two scales on MMPI, GSR, and Cattell 16 PF tests were significant in differentiating underachievers from normal achievers. Underachievers are less concerned with social approval, more nonconforming, more tense, more emotionally unstable, have lower frustration tolerance, more trustful and more warm than controls. Suggests that GSR initial resistance and that obtaining after 5 minutes' rest are reliable predictors of under achievement.

100. Young, C. W. and Eastabrooks, G. H. Non-intellectual factors related to scholastic achievement. Psychol. Bull., 1934, 31, 735-736.

A studiousness index (SI) was computed on 582 students using relative GPA with I.Q. held constant. Using the highest and lowest 100 SI's, an analysis of Personal Inventories, Strong VIB yielded weighted items to differentiate high and low groups. Cross-validation on 275 students gave correlation with grades of .27 for Inventory and .34 for the VIB. Studiousness not related to intelligence. Studious persons tend to be idealistic, liberal on social and economic questions, conservative on moral ones. In self-ratings, they percieve themselves as cautious, conscientious, industrious, self-conscious, indifferent to pleasure, selfish (but self-sacrificing on principle), and self-sufficient. In general, they resemble, according to the authors, the usual picture of the introvert.

# A RATIO APPROACH TO THE MEASUREMENT OF OVERACHIEVEMENT-UNDERACHIEVEMENT

G. Douglas Mayo

Naval Air Technical Training Command

After hearing Dr. Thorndike enumerate the pitfalls associated with overachievement-underachievement and the relatively remote possibility of dealing with anything more substantial than statistical artifacts, one is more inclined to become a critic of the work of others than he is to attempt to do anything constructive on his own in this area. One must be reasonably bold in order to propose methodology, statistical or otherwise, as Dr. Thorndike and Dr. DuBois have done. To describe empirical results in the area of overachievement-underachievement, or even research that is underway, is unquestionably foolhardy. And yet this is precisely what I am here to do. Moreover, since it is doubtful that I can get into much more trouble by citing several studies than by citing a single study in the area of overachievement-underachievement, I shall mention more than one.

I should like to view the problem essentially in this frame of reference: it may well be that the overachievement-underachievement problem is a pseudo research area in the sense that the unpredicted variance in school grades (or whatever criterion of achievement is used) could more appropriately be called underachievement on the part of the psychologist in his effort to predict student achievement than underachievement on the part of the student. I share the hopes of those who take the point of view that the real problem is the development of predictors that will account for virtually all of the predictable variance in the criterion. I share their hope that the state of the art will eventually permit this and leave no variance that could be identified as overachievement-underachievement variance. Obviously the state of the art does not permit this at the present time. It does, however, permit us to do some fairly interesting things with the variance in the criterion which as yet we are unable to predict with aptitude tests. Also of great interest is the fact that under certain conditions variance not associated with measured aptitude in a sample of behavior may be used on the predictor side of the ledger, as we shall see in a moment.

With respect to tools that are available to us, Dr. DuBois has shown that a properly constructed ratio, such as the ratio of the actual grade made in a course of instruction, to the grade predicted by the aptitude scores, is the equivalent of the residual which remains when the variance which the aptitude test has in common with the grades made in the course is partialed out. The practical significance of this lies in the fact that we do not have to compute overachievement-underachievement scores in our research work involving this variable. We may do any sort of correlational analysis we wish with this residual, that is, with the course grade, with aptitude partialed out, and then when we wish to apply this information to individuals in the school setting we may compute the ratio of school grades achieved to school grade predicted for an individual from aptitude scores with the assurance that the resulting scores have essentially the same validity as did the variable expressed as a residual. In other words, the ratio is a practical equivalent of the residual and permits identifying a score for each individual, whereas the residual procedure does not. Perhaps some concrete examples of studies involving the ratio and the residual would be in order at this point to illustrate some of the possibilities mentioned in a general way a moment ago.

The rationale of the first study we shall mention was as follows: a relatively small sample of performance in a setting similar to one in which it is desired to predict subsequent success, frequently will predict performance in the latter situation better than aptitude tests will. The total variance in the earlier sample of performance may be divided by partial correlation procedures into two parts, the part that is completely overlapped or perfectly correlated with aptitude on the one hand and the part that is linearly uncorrelated with aptitude on the other.

Now clearly this latter part or residual contains all variance that was contained in the initial sample of performance other than that associated with aptitude as measured by the aptitude test used. Doubtless this includes aptitude <u>not</u> measured by the particular tests used, previous achievement in the area of the sample of performance, manifestations of effort exerted by individuals in the sample of performance, and numerous other variance contributors, several of which have been enumerated by Dr. Thorndike.

There are two matters involved here. The first is the probability that we can improve prediction by using the brief sample of performance, and if we wish to, by means of the ratio approach already mentioned, we may express the quality

of this performance in two scores rather than one, namely an aptitude score and the part of the sample of earlier performance that is unrelated to the aptitude measure. Why would we want to do this? Well, not everyone does want to do it. But among the ones who do want to are two general classes of people. The first, whom we shall mention only briefly, are administrative people who are accustomed to thinking in terms of scores on aptitude tests with which they are familiar. They are pleased to improve prediction by means of an additional measure, which correlates zero with aptitude and therefore adds unique variance in the prediction situation. They could use the score on the brief sample of performance, but they would prefer not to have aptitude variance contained to some unknown degree in the score representing the sample of performance. The score resulting from the ratio can, of course, be expressed in standard scores or percentile scores, utilizing the same metric as that of the aptitude test. The second group that might want to divide the variance in the sample of performance into the two parts referred to above are those who would like to ask certain questions about the nature of the sample of performance after the variance associated with aptitude has been removed. They view the pervasive character of aptitude - its tendency to relate to most other variables, to various degrees - as a contaminating influence which may obscure the relationships in which they are interested. Some of the questions one might wish to ask are: Can a sample of performance be found which will add appreciably to the prediction currently available in aptitude tests in predicting subsequent performance? To what extent is the sample of performance with aptitude removed related to effort exerted during the period of the performance? If we may be permitted to call this residual or ratio "overachievement-underachievement" (in quotes), how stable is it over time? Is it largely random error or will overachievement-underachievement in a sample of performance predict overachievement-underachievement in subsequent performance? Since it appears that evidence concerning these questions would be preferable to speculation, data will be presented on these points.

The data were collected in the Naval Air Technical Training Command. In the first study the early sample of performance was provided by an airman preparatory school which prepares men, who have just completed recruit training, for training in the 12 basic occupational specialties for enlisted personnel in Naval Aviation. The school curriculum includes such basic subjects as physics, mathematics, handtools, layout, and occupational information concerning Naval Aviation job specialties. The course is six weeks in length. In this study the measure of "overachievement-underachievement" (still

in quotes) was the ratio of the grade actually made in the airman preparatory course to the grade predicted by a battery of aptitude tests given routinely about three months earlier at the recruit training command.

Students who made higher grades on the course than predicted by their aptitude scores received over-achievement-underachievement scores greater than one and students making lower grades than predicted received scores less than one. The aptitude score used to obtain the predicted grades in the airman preparatory school was the sum of three tests, the Navy General Classification Test, which is essentially a verbal intelligence test; the Navy Arithmetic Test which includes both computational and reasoning items; and the Navy Mechanical Test which is a picture and question type test similar to the Bennett Mechanical Comprehension Test. The correlation between this composite aptitude measure and airman preparatory school grades was .76.

Both the composite aptitude scores and the overachievement-underachievement scores were then correlated with grades made in the basic schools which train for the 12 Naval Aviation occupational specialties. These correlations are shown together with means and standard deviations in Table 1. It is noted that the correlation between basic aviation technical school grades and aptitude scores tend to cluster around a median of about .57. The median of the 12 correlations involving over-achievement-underachievement is about eleven correlation points lower at .46. The multiple correlations between basic technical school grades on the one hand and the aptitude battery and over-achievement-underachievement on the other are also shown in the table. The median of these twelve multiple R's is .79. This is excellent prediction in terms of our usual expectations in psychological work. But of course no credit for this is due to the ratio or residual approach since it is the preparatory school grades that correlate to this extent with the performance in the basic schools. All the ratio procedure or residual procedure does is divide the variance in the sample of behavior represented by grades into two parts, as we have said before. Hence, it can contribute nothing to prediction, as such, that is not already in the preparatory school grade. However, we have demonstrated that in a practical situation it is quite possible to find a sample of performance which will add substantially to a battery of aptitude tests in the prediction of subsequent performance.

We now consider the question of any evidence we may bring to bear on the question of the nature of the over-

achievement-underachievement scores as derived by the ratio or residual procedure. One approach to this question is through the use of peer rating on effort expended by basic school students to master the content of the courses they are taking. In collecting the peer ratings, instructional groups of about 15 to 25 men were asked to nominate and rank the three students in their section who were trying the hardest to master the course, also nominate and rank the three students who were trying the least hard. Nominations were weighted by rank order, summed algebraically and converted to standard scores. It will be noted that we are attempting to predict effort expended at a later point in time in the basic schools from the overachievement-underachievement ratio in the airman preparatory school, the sample of performance on which the ratio was based having been observed several weeks earlier. Peer rating data were collected in four of the basic schools as shown in Table 2. With the exception of the electronics technician school the correlations tend to be in the .33 to .50 range. The correlation in the case of the electronics technician school is substantially lower at .17, correctable for restriction in range to .24 (if one wished to make such a correction), but still statistically significant with an N of 209.

Doubtless we could agree at a descriptive level that the overachievement-underachievement ratio tends to predict peer ratings on effort in certain basic schools. Just what this means or why this is the case might be quite another matter. Peer ratings have little or no more immunity to criticism than does overachievement-underachievement.

One of the questions which arises in this context is the extent to which a generalized favorable or unfavorable impression of a man on the part of his classmates influences peer ratings on effort. For example, would peer ratings on effort correlate just as well with an objective measure of intelligence as they did with the overachievement-underachievement ratio which purports to reflect effort to some degree? Would the correlation between peer ratings on effort and peer ratings on intelligence be approximately as high as the reliabilities of the two sets of peer ratings would permit? In other words, are peer ratings on two favorable attributes such as intelligence and effort perfectly correlated except for attenuation attributable to unreliability of the peer rating measures? Again, we can present some evidence on these points.

In one of the basic schools already mentioned, the engine mechanic school, scores on the Navy General Classifi-

cation Test, which has already been described as essentially a verbal intelligence test, were obtained. At the time peer ratings on effort were obtained, peer ratings on intelligence also were collected. As previously noted overachievement-underachievement ratio scores were available. The intercorrelations among these variables are shown in Table 3. It is noted with respect to the question of whether peer ratings on effort would correlate as well with intelligence test scores as with the overachievement-underachievement ratio, that this is not the case. The correlation of peer ratings on effort with the overachievement-underachievement ratio is .40 while the correlation with intelligence test scores is .29. With respect to the question concerning whether the correlation between the two sets of peer ratings is as high as the reliabilities would permit, the reliabilities of the peer ratings were estimated by a procedure which involved dividing the raters into two groups and correlating the resulting arrays of scores. The reliabilities of peer ratings on effort and peer ratings on intelligence were .84 and .85 respectively. Since the maximum correlation which may be obtained between two measures is the square root of the product of the reliabilities of the two measures, a correlation of about .84 would be possible in this case. The observed correlation of .66 as shown in Table 3 is high, but substantially less than .84.

Further evidence on this point results when the variance associated with peer ratings on intelligence is partialed out. The remaining correlation between the overachievement-underachievement ratio and peer ratings on effort is still statistically significant, although only at the .05 level.

This all sounds very good, but as usual there is at least one insect in the ointment. For some reason, which I shall need some assistance to explain, the overachievement-underachievement ratio correlates somewhat higher with peer ratings on intelligence than it does with peer ratings on effort, .46 as opposed to .40. This suggests that something is wrong, but it is doubtful that my speculating on what it is would be very enlightening.

A final question that was raised was the matter of the stability of the overachievement-underachievement ratio or residual over time. Information on this topic was

obtained in connection with a larger study which was concerned primarily with another problem. This study involved 196 students in the basic school for aviation structural mechanics. The part of the study in which we are interested here correlated the overachievement-underachievementment residual, that is, the airman preparatory school grade with the three Navy aptitude tests previously mentioned partialed out, with the overachievement-underachievement residual in two units of the school for aviation structural mechanics. The first unit was entered approximately one month after completion of the airman preparatory school. This unit consisted of instruction in sheet metal and was five weeks in length. The second unit involved instruction in welding and followed the sheet metal unit. It was two weeks in length.

As shown in Table 4 the overachievement-underachievement residual in airman preparatory school predicts the overachievement-underachievement residual in the sheet metal unit to the extent of .49 and the overachievement-underachievement residual in the welding unit .35. The two units of instruction residuals correlate .53. This seems to indicate that the overachievement-underachievement residual has some stability over time, that regardless of other faults it may have, its existence is not invariably dependent upon random error.

By way of summary, to this qualified kind word for overachievement-underachievement, I would add that overachievement-underachievement, as measured by the ratio or the residual, may serve a useful function as a predictor of subsequent performance and that it relates to effort expended in the school environment as measured by peer ratings. These points suggest that it may be premature to write off the area of overachievement-underachievement as one in which no constructive work can be accomplished.

The problems concerning overachievement-underachievement enumerated by Dr. Thorndike are, of course, very real and clearly must be taken into account, but it is possible that they may not exert quite as much influence on carefully collected actual data as one might suppose when they are considered in the abstract. Perhaps this one cheerful note would be an appropriate one on which to conclude my remarks.

Table 1

Correlation Between Basic Aviation Technical School Grades and Overachievement-Underachievement Ratio and Aptitude

| Basic Aviation Technical School Grades | N | Overachievement-Underachievement Ratio | | | | Aptitude | | | | Multiple R |
|---|---|---|---|---|---|---|---|---|---|---|
| | | r | $r_c^*$ | M | S.D. | r | $r_c^*$ | M | S.D. | |
| Boatswain | 250 | .43 | .47 | .975 | .059 | .50 | .55 | 157.64 | 16.90 | .78 |
| Aircontrolman | 249 | .23 | .30 | 1.006 | .050 | .37 | .41 | 173.78 | 17.32 | .62 |
| Engine Mechanic | 250 | .46 | .49 | .994 | .061 | .62 | .65 | 151.68 | 17.82 | .81 |
| Electrician | 249 | .48 | .45 | .995 | .068 | .60 | .62 | 152.73 | 18.49 | .77 |
| Aerographer | 249 | .21 | .27 | 1.027 | .051 | .56 | .60 | 171.26 | 17.37 | .83 |
| Storekeeper | 220 | .41 | .46 | 1.009 | .057 | .30 | .34 | 160.81 | 16.82 | .60 |
| Structural Mechanic | 236 | .48 | .49 | .999 | .065 | .62 | .63 | 151.71 | 18.60 | .83 |
| Ordnanceman | 249 | .49 | .50 | .973 | .063 | .59 | .64 | 147.33 | 17.38 | .81 |
| Electronics Technician | 250 | .55 | .58 | 1.006 | .061 | .61 | .64 | 163.38 | 18.02 | .88 |
| Photographer | 249 | .31 | .36 | 1.008 | .056 | .31 | .34 | 161.44 | 17.48 | .53 |
| Parachute Rigger | 60 | .60 | .60 | .989 | .065 | .26 | .39 | 161.00 | 12.61 | .75 |
| Training Deviceman | 246 | .24 | .33 | 1.022 | .047 | .50 | .55 | 176.44 | 17.28 | .88 |

*Corrected for restriction in range using standard deviation of an Airman Preparatory School Sample of 1000.

Table 2

Correlation of Overachievement-Underachievement Ratio with Peer Ratings on Effort

| Aviation School | N | r | $r_c^*$ |
|---|---|---|---|
| Engine Mechanic | 166 | .40 | .41 |
| Structural Mechanic | 167 | .33 | .44 |
| Electronics Technician | 209 | .17 | .24 |
| Training Deviceman | 266 | .39 | .50 |

*Corrected for restriction on Overachievement-Underachievement using standard deviation of an Airman Preparatory School Sample of 1000.

Table 3

Intercorrelations among Two Peer Ratings, Overachievement-Underachievement Ratio, and Intelligence (N = 166)

| | Peer Ratings on Intelligence | Peer Ratings on Effort | Intelligence (Navy GCT) |
|---|---|---|---|
| Peer Ratings on Effort | .66 | | |
| Intelligence (Navy GCT) | .46 | .29 | |
| Over-Underachievement | .46 | .40 | .10 |

Table 4

Intercorrelations among Overachievement-Underachievement Residuals Derived in Three Instructional Situations at Different Points in Time (N = 196)

| Over-Underachievement Measures | Sheet Metal Unit Residual | Welding Unit Residual |
|---|---|---|
| Airman Preparatory Residual | .49 | .35 |
| Sheet Metal Unit Residual | | .53 |